TEA y TDA: Autismo y Topología

Predicción y diagnóstico de Autismo con Análisis Topológico de Datos
utilizando Resonancias Magnéticas Funcionales

Federico Tomás Poncio
Director: Claudio Delrieux
Co-Director: Emmanuel Iarussi

Trabajo de tesis presentado para el título de
Magíster en Data Mining

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Argentina
Septiembre 2021

Índice

Agradezco a mis directores, Claudio y Emmanuel,
por su compromiso y guía en este proyecto.
Fue un gusto trabajar con ustedes.

1. Introducción

1.1. Motivación y Problema

Las condiciones encapsuladas en los Trastornos del Espectro Autista (TEA) han recibido un interés creciente debido al aumento reciente de diagnósticos de autismo, pasando de 1 de cada 150 niñes de hasta 17 años en los 2000, a 1 de cada 54 en 2016 en los Estados Unidos[1]. La temprana detección de los TEA resulta crucial para el éxito de sus intervenciones, lo que motiva la búsqueda de un método que no dependa de que el niñe presente señales de autismo como, por ejemplo, dificultades en la comunicación o el desarrollo del lenguaje. Es en este respecto que cobran relevancia los conectomas. Los conectomas son representaciones de las conexiones nerviosas de sujetos a distintas escalas: desde regiones de interés en el cerebro hasta el mapeo de las conexiones a nivel neuronal. Partiendo del éxito en el mapeo del conectoma de los nemátodos y conectomas de la retina de un ratón, se destaca el trabajo del Human Connectome Project[2], que busca mapear el conectoma humano. De estas ideas surgió en el 2018 la competencia IMPAC: IMaging-PsychiAtry Challenge - Predicting Autism. Allí se disponibilizó un conjunto de entrenamiento y uno de testeo de aproximadamente mil muestras cada uno, de fMRIs de sujetos con y sin TEAs (50/50) con el objetivo de utilizar estos conectomas en la predicción. La competencia terminó, pero los datos están ahora públicamente disponibles.

Asimismo, existen un conjunto de ideas y técnicas que comprenden el Análisis Topológico de datos (TDA por sus siglas en inglés) que se destacan por su idoneidad para resumir y analizar aspectos inherentes a la forma de los datos. Esto provee la motivación de utilizar a la topología en el análisis de los conectomas. El TDA deriva de la topología su capacidad para extraer información sobre los datos que sea invariante ante traslaciones y deformaciones continuas (un cambio de escala o una deformación de los datos puede tener gran impacto en los resultados de algunos algoritmos, pero topológicamente siguen siendo el mismo objeto). Asimismo, el TDA ha provisto herramientas que son estables ante la presencia de ruido. Esto sumado a la hipótesis de que los datos provienen de una variedad [3] predeterminada, permite la incorporación de nociones de significatividad de los valores hallados (Ver Michel (2015) y Berry (2020)). En otras palabras, puede asumirse que los

[1] Ver `https://www.cdc.gov/ncbddd/autism/data.html`

[2] `http://www.humanconnectomeproject.org/`

[3] Una variedad (o *manifold* en inglés) de dimensión n es un espacio que localmente 'se parece' a $\mathbb{R}^n$. Que los datos provengan de una variedad desconocida no es un supuesto ajeno al análisis de datos: por ejemplo, en una regresión lineal se asume que los datos provienen de una línea o un hiperplano (que son variedades), con un término de error Gaussiano añadido.

datos provienen de una variedad desconocida $\mathcal{M}$, más un término de ruido; y las características topológicas relevadas tienen poca sensibilidad a estos errores.

Se puede ver en la literatura reciente un interés por utilizar la topología de los conectomas como features para diagnosticar afecciones de origen neurológico. El objetivo de este trabajo es continuar este camino concentrándose en particular en los Trastornos del espectro Autista; y ver si los resúmenes topológicos de los conectomas aportan al poder de diagnóstico actual. De encontrar un aporte positivo abriría la puerta a la conclusión de que el autismo afecta no solamente a los niveles de conectividad cerebral sino también a la conformación topológica del conectoma en sí.

En este trabajo utilizaremos el dataset de la competencia IMPAC para comparar la performance de algoritmos de predicción que utilizan los niveles de conectividad con aquellos que también incorporan resúmenes de la topología de los conectomas. Exploraremos el poder predictivo de las características topológicas por sí mismas, las *features* derivadas de ellas, y los efectos de los diferentes caminos posibles en su preprocesamiento. En cualquier caso, se intentará responder por la positiva o la negativa la utilidad de la incorporación de estas nuevas técnicas en los pipelines estándares de diagnóstico neurológico. En el camino se explorarán las aristas del problema y los aportes que la topología puede tener en este área.

1.2. Organización del trabajo

Este trabajo se organiza de la siguiente manera: comenzando en la próxima sección 2, se introducirá el estado del Arte y el tratamiento actual de los problemas de diagnóstico de los TEA. En la sección 3 se presenta una introducción y un tratamiento más en profundidad de la teoría de la homología que compete a este trabajo, y luego en la sección 4 se expone la homología en su formalismo matemático estándar. En la sección 5 se explora el dataset del IMPAC y demás conjuntos de datos utilizados en este trabajo, para luego describir en la sección 6 los métodos utilizados en la exploración y predicción, así como las métricas utilizadas para evaluar los resultados. Luego en la sección 7 se presentan los resultados de los experimentos llevados a cabo, concluyendo en la sección 8 con la discusión final del trabajo.

2. Estado del Arte

Los trastornos del espectro autista son un tipo de trastornos del neurodesarrollo que se caracterizan por dificultades en la sociabilización, comunicación, y presencia de intereses restringidos y conductas estereotipadas. Inicialmente en los 60's y 70's se consideró al Autismo como un diagnóstico único, para luego complejizarse la caracterización e incluir subtipos de este trastorno. Lo que conocemos hoy en día como Trastornos del Espectro Autista comprende al Autismo, el Asperger, y el Trastorno generalizado del Desarrollo no especificado. De los diagnósticos en niñes de 8 años, se estima que 14.7 cada 1,000 niñes poseía un trastorno autista en 2010, comparado con 11.3 en 2008, 9.0 en 2006, 6.6 en el 2002, y 6.7 en el 2000 (Christensen et al., 2018). En (Hong et al., 2019, 2), remarcan: "Despite ample research on the brain basis of autism, a neurobiological framework to consolidate the co-occurrence and interplay of low- and high-level functional abnormalities remains to be established" [A pesar de la amplia investigación sobre las bases neurológicas del autismo, no se ha podido establecer aún un marco neurobiológico que consolide la co-ocurrencia e interacción de las abnormalidades funcionales de alto y bajo nivel]. Allí, comentan también que al momento hay poco acuerdo sobre los hallazgos relativos al diagnóstico del autismo de fMRIs.

En el modelado de datos de fMRI y en lo que compete a este trabajo, hay tres líneas de estudio presentes en la literatura a tener en cuenta. La primera se enfoca en el modelado de conectomas como grafos, con sus problemáticas asociadas. Luego la segunda refiere a la utilidad en particular del TDA para estudiar este tipo de datos, y la tercera específicamente a la predicción de autismo, tanto con TDA como sin.

En Bassett and Sporns (2017) tratan particularmente un tema que llaman "network Neuroscience": la interesección entre la neurociencia y el estudio de los grafos. Presenta ideas acerca de cómo modelar estas series temporales por regiones como grafos, comentando la necesidad de incorporar las matrices de correlación en el modelado. La discusión está centrada principalmente en técnicas y medidas estándares de trabajo de grafos (conectividad, por ejemplo), y la mención de la homología persistente es breve. Un aporte interesante del trabajo puede verse en la figura 1: puede apreciarse aquí el lugar que ocupa la fuente de datos de este trabajo, en la intersección de la sección de fMRI con la de redes de conectomas.

En una línea similar, pero concentrándose más en la distinción de sujetos particulares y no en la predicción de condiciones o el modelado del conectoma en sí, Venkatesh et al. (2019) proponen utilizar una distancia geodésica para computar la cercanía entre dos matrices de conectividad, y arguyen que se ajusta mejor al problema porque toma en cuenta la geometría particular del espacio

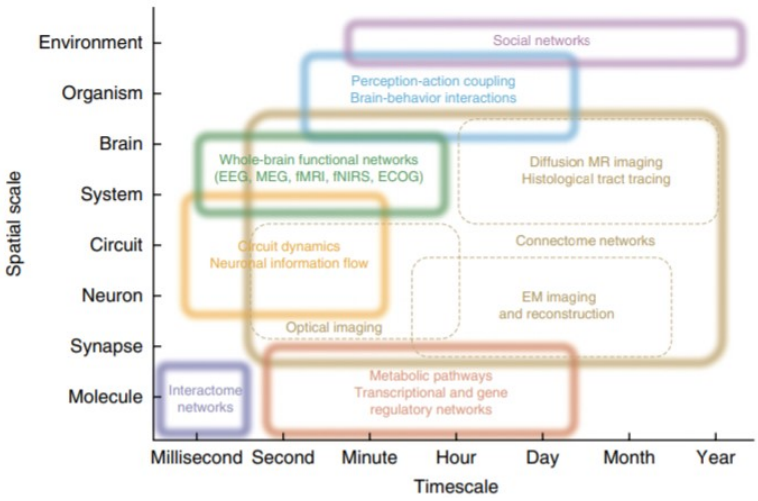

Figura 1: Organización de los distintos estudios neurológicos según los intervalos temporales que cubren y su escala espacial (Bassett and Sporns, 2017, 3)

.

de matrices de correlación (por ejemplo, porque restar matrices de correlación no siempre resulta en otra matriz de correlación, requiriendo un trato especial para mantenerse siempre dentro del espacio apropiado). Este enfoque es en parte retomado por este trabajo, al incorporar la descomposición tangencial (entre otras) para el cálculo de las matrices de covarianza. La maquinaria matemática de estos cómputos puede encontrarse en Pennec et al. (2006), donde exponen con detalle la relación entre estos conceptos.

En Dadi et al. (2019), les autores recorren el estado del arte en modelos predictivos a partir de fMRIs, con el fin de estandarizar las prácticas científicas del área y determinar un baseline de comparación para los varios modelos posibles. Cabe destacar que no incluye modelos que utilicen TDA en sus predicciones. Utiliza distintos datasets para diagnosticar distintas afecciones; entre ellas adicciones, esquizofrenia, alzheimer, y autismo. En particular con respecto al autismo encuentran los mejores resultados usando modelos logísticos con penalización $L2$ y llegan en sus máximos resultados a un 75 % de accuracy en la predicción, y un 69 % en media. También se ve una pequeña mejoría en la clasificación con el uso de la descomposición tangencial de las matrices de covarianza en contraste con el uso de matrices de correlación o correlación parcial.

Sobre el uso del TDA en el ámbito neurocientífico en general, en Ellis et al. (2019) evalúan la plausibilidad de utilizar Análisis Topológico para detectar la estructura de datos de tipo fMRI, y encuentran que el TDA puede detectar las estructuras presentes en datos simulados en Python.

Utilizando el TDA, en Stolz et al. (2020) comparan las homologías de pacientes con esquizofrenia con sujetos de control en fMRIs de reposo, y utiliza transformaciones como las imágenes de persistencia o *landscapes* de persistencia como datos de entrada para algoritmos de K-medias con el fin de facilitar esta comparación. De todas formas, tienen un objetivo más descriptivo que predictivo, pero los pasos a seguir en la detección de diferencias en la conectividad funcional para el diagnóstico de esquizofrenia y de autismo son prácticamente los mismos, entonces los aprendizajes o limitaciones de estas técnicas en relación a la primera pueden ser útiles para nuestra tarea. En general en este trabajo no se encuentran grandes resultados de detección de esquizofrenia, ya que la separación de clases no fue significativa salvo en algunos grupos.

Tenemos también el trabajo de Caputi et al. (2021) donde analizan de forma más general el campo de posibilidades del TDA para el análisis de conectividad funcional cerebral. En este artículo les autores remarcan que en la mayoría de sus experimentos las features de topológicas de la conectividad cerebral tuvieron una performance casi aleatoria en el diagnóstico de esquizofrenia. Sugieren que el TDA puede encontrar su nicho en aquellas aplicaciones donde los grafos de conectividad no son directamente computables;por ejemplo, en aquellas técnicas de medición de actividad cerebral que no mantienen el mismo punto de referencia para cada sujeto y no permiten estandarizar las medidas en una misma matriz de conectividad. Un punto importante a remarcar es que exploran también los usos de homología *dirigida*, y no solamente de la homología como se trata en este trabajo. La homología dirigida es conceptualmente similar a la homología persistente, con la diferencia de que toma como input un grafo dirigido, y computa las homologías con esta mayor complejidad incluida.

Por otro lado, en la búsqueda de la predicción específica de los trastornos del espectro autista, Byrge and Kennedy (2020) utilizan casi dos horas de fMRIs basados en eventos por individuo y no llegan a resultados prometedores. Menciona también la dificultad en la predicción del autismo en otros trabajos. De todas formas, sus métodos de predicción utilizan modelos lineales generalizados sobre las series temporales y luego modelos logísticos, pero no utilizan métodos que extraigan características topológicas de los datos.

Rathore y sus colegas (Rathore et al., 2019) escribieron un artículo que tiene un objetivo similar al de este trabajo, pero utilizan otro dataset abiertos de resonancias magnéticas de autismo llamado ABIDE[4] (que es el mismo que utilizan Dadi et al. (2019)). Sus resultados advierten sobre un límite superior al poder predictivo de estas técnicas sobre la predicción del autismo. Analizan

[4]Puede encontrarse más información sobre este dataset en `http://preprocessed-connectomes-project.org/abide/`

también técnicas de homología persistente en comparación con matrices de conectividad y llegan a un 69.2 % de aciertos en la predicción de autismo con estos datos. Advierten también que la mejoría encontrada al agregar datos topológicos a sus modelos no siempre es estadísticamente significativa. En este trabajo vamos a retomar enfoques como los expuestos en su artículo y explorar en mayor profundidad las capacidades predictivas de los fMRIs y el aporte de las técnicas de homología persistente en el diagnóstico.

3. Análisis Topológico y homología persistente

A continuación se hace un recorrido por la teoría de la homología y, en general, el marco teórico matemático que subyace a este trabajo. Esta sección está organizad para servir también como una introducción al concepto de la homología persistente, y se ordenó según su nivel de abstracción y formalismo. La sección 3.1 presenta una exposición informal del tema que busca acercar estas herramientas a quienes les interese, sin entrar en los detalles más formales y difíciles de la teoría. La sección 3.2 continúa en este camino aumentando el nivel de formalismo involucrado y proveyendo más contexto para quienes les interese acercarse más al tópico. Las secciones posteriores continúan presentando temas relevantes a este trabajo relacionados a la homología. Luego, la sección 4 presenta a la homología simplicial en su formalismo estándar, junto con preliminares matemáticas en caso que se desee ahondar más en este contexto.

Para más información y fuentes referentes a esta sección se recomienda los textos de Edelsbrunner and Harer (2010) y Aktas and Fatmaoui (2019), así como las notas de curso de Vidit Nanda que pueden encontrarse en `http://people.maths.ox.ac.uk/nanda/cat/TDANotes.pdf`, y la presentación de Frédéric Chazal y Bertrand Michel en `https://geometrica.saclay.inria.fr/team/Fred.Chazal/Barcelona2016/slides/PersistenceForTDA.pdf`.

3.1. Introducción

El análisis topológico de datos es una rama de estudio que surgió recientemente y busca utilizar conceptos y herramientas de la topología y aplicarlos al análisis de conjuntos de datos. La topología estudia principalmente lo que se conocen como *invariantes topológicos*, que son atributos que caracterizan a objetos con respecto a su "forma". El ejemplo arquetípico de esto es el número de agujeros que algo tiene: si nos imaginamos un retazo de tela elástica con un círculo cortado en el medio, independientemente de cuánto estiremos o comprimamos el retazo, el círculo va a seguir

estando ahí y va a seguir siendo solo uno. En este ejemplo, la cantidad de agujeros es un invariante topológico, y lo que lo diferencia de otros invariantes (no topológicos) es que son invariantes *ante deformaciones continuas del objeto*; léase: sin cortarlo ni romperlo, sólo estirándolo.

Así como saber el valor promedio y la varianza de un conjunto de observaciones puede ayudarnos a entender cómo es "su forma", comprender los invariantes nos da herramientas formales para describir a un objeto y diferenciarlo de otros. En el caso de este trabajo, por ejemplo, nos permitiría diferenciar a un conectoma de otro.

Una de las grandes herramientas que posee el Análisis topológico de datos -TDA de ahora en más- para capturar los invariantes se conoce como **homología persistente**, que es a su vez una aplicación particular de la teoría de la homología en general. En la homología se buscan los "grupos de homología", donde cada grupo nos permite saber cuántos agujeros tiene nuestro espacio topológico en cada dimensión. La dimensión más fácil de asequir es la primera: el círculo cortado del retazo de tela es un agujero unidimensional. Esta información va a estar presente en el grupo de homología de dimensión 1, escrito H_1. El grupo de homología de dimensión 2 (H_2) captura agujeros bidimensionales. Un ejemplo de esto es una pelota: la pelota divide al espacio tridimensional en dos sectores, uno interno y otro externo; el interno siendo el agujero bidimensional. El razonamiento puede extenderse para cualquier grupo de homología H_k. Con respecto a H_0, un agujero de dimensión cero sería un espacio entre componentes, entonces la homología de dimensión cero captura los distintos *componentes conexas* del espacio.

A modo de ejemplo pensemos en una pelota nuevamente. La pelota tiene solo una componente conexa (que sería el plástico que la conforma); no tiene ningún agujero unidimensional (porque sino se pincharía), y tiene un agujero bidimensional, encapsulado por el plástico. Las dimensiones entonces de nuestros grupos de homología serían 1, 0, y 1, para H_0, H_1, H_2, respectivamente.

La homología en sí es un poco más complicada que el ejemplo de recién, por eso fue más conveniente hablar de la dimensión de los grupos en vez de los grupos en sí. Las dimensiones de cada grupo de homología se llaman **números de Betti**, y se escriben con la letra griega beta (β). En el ejemplo de la pelota entonces, lo correcto sería decir que $\beta_0 = 1$, $\beta_1 = 0$, y $\beta_2 = 1$.

Lo importante para entender en términos de la intuición es que la homología de grado k captura información sobre los agujeros de dimensión k que tiene un espacio determinado. H_0 mide la cantidad de componentes conexas, H_1 la cantidad de loops o agujeros unidimensionales, y H_2 la cantidad de vacíos o agujeros bidimensionales. Evelyn Lamb dice:

> "If you can put it on a necklace, it has a one-dimensional hole. If you can fill it with toothpaste, it has a two-dimensional hole. For holes of higher dimensions, you're on your own"[Si se puede poner en un collar, tiene un agujero unidimensional. Si se puede llenar de pasta de dientes, tiene un agujero de dos dimensiones. Para agujeros de mayores dimensiones, estás por cuenta tuya] [5].

Con la base de la homología asentada, necesitamos ahora complejizar un poco el panorama. La dificultad va a provenir de que en general cuando se trabaja con datos se utilizan finitas muestras -con ruido- para estudiar un fenómeno mayor que los datos; ya sea una muestra de personas de quienes sabemos sus ingresos y sus edades para hacer una regresión, o mediciones de precipitaciones y humedad para predecir el clima al otro día, entre muchos otros ejemplos. Incluso en el ejemplo de la pelota sería inusual tener la pelota entera, sino que se tendrían puntos de su superficie como referencia que, nuevamente, pueden contener ruido. Resulta entonces importante tener alguna forma de reconstruir el objeto original a partir de muestras finitas.

Una forma de reconstruir el espacio topológico original (de donde teóricamente se tomaron las muestras) es creando un grafo a partir de las observaciones: se fija un radio $\varepsilon > 0$, y a cada punto se le superpone una bola de radio ε, y cuando la intersección entre dos bolas es no vacía, a los puntos se los une con un eje. Esto es equivalente también a unir con un eje a todos los puntos que estén a distancia menor que ε. Este grafo va a tener un cierto número de componentes conexos, loops, etc, y puede entonces computarse la homología del grafo. Puede verse una representación gráfica de esto en la figura 2.

En la descripción anterior resulta clave el valor de ε que se utiliza, ya que determina qué puntos se unen y cuáles no, por ende afecta a las homologías que se computan del grafo resultante. Así nace entonces la *homología persistente*, con la idea de probar todos los ε posibles y ver qué características *persisten*. La intuición sería que cualquier agujero que represente una característica real de los datos "existiría"por más tiempo que aquellas surgidas por artefacto del ruido presente.

La homología persistente se representa gráficamente de dos formas: diagramas de persistencias, y diagramas de barras. En ambos casos la idea es la misma. Primero se fija una dimensión, por ejemplo $k = 0$ para ver los componentes conexos. Poniendo el foco entonces en H_0, se computa ese grupo de homología para cada ε entre 0 e infinito. Cuando en el grafo resultante se ve una componente conexa que en el ε anterior no existía, se dice que esa componente nace. El ε donde nace esa i-ésima componente conexa se guarda bajo la variable b_i (por birth en inglés). Asimismo,

[5] En: https://blogs.scientificamerican.com/roots-of-unity/what-we-talk-about-when-we-talk-about-holes/

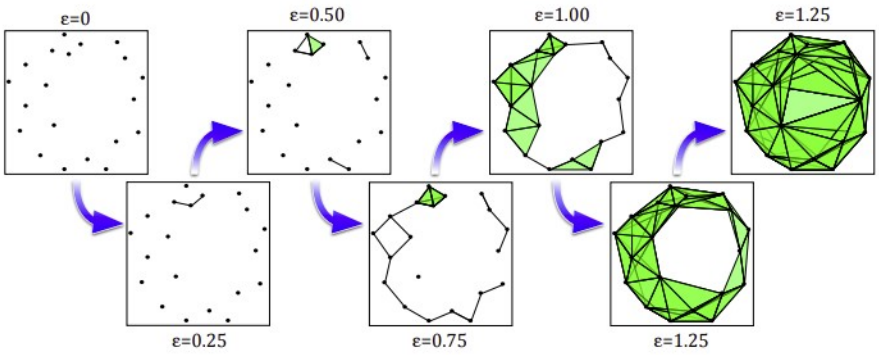

Figura 2: Ejemplo de la construcción de un grafo a partir de puntos en el espacio a distintos niveles de distancia ε. A cada grafo puede luego computársele la homología, y analizar cómo cambia según el ε. En Medina and Doerge (2016)

cuando esa componente es absorbida en otra mayor o deja de existir, el ε que corresponde se guarda con nombre d_i (por death en inglés). Al terminar de recorrer todos los ε posibles, se obtienen una serie de puntos de la forma (b_i, d_i), que corresponden con las componentes conexas que nacieron y murieron al recorrer las homologías de orden 0. Para crear el diagrama de persistencia de orden 0 con estos datos, se hace un gráfico de puntos colocando en el eje de las abcisas los nacimientos y en el eje de las ordenadas las muertes. De forma similar en los diagramas de barras se traza una línea por cada componente que comienza en b_i y termina en d_i, y se "apilan"las lineas de las distintas componentes por orden de aparición.Este mismo procedimiento puede repetirse para las homologías de cualquier orden, esta vez midiendo los agujeros de dimensiones mayores. Podemos ver un ejemplo de esto en la siguiente figura[6]:

3.2. TDA y homología persistente

3.2.1. Intuición de la Homología

El objetivo de esta sección es dar una presentación más formal del concepto de homología que sirva como contexto teórico de los experimentos conducidos en este trabajo. Una pregunta que puede pensarse que guía esta sección es: ¿cómo puede hacerse para contar los agujeros de un espacio topológico? Para responder vamos a adentrarnos más formalmente en la homología. Quien desee revisar los conceptos de topología, espacio topológico, espacio vectorial, grupos, grupos de cociente, que también son preliminares a la homología, puede dirigirse a la sección 4.1.

[6]En https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7390646/ REF

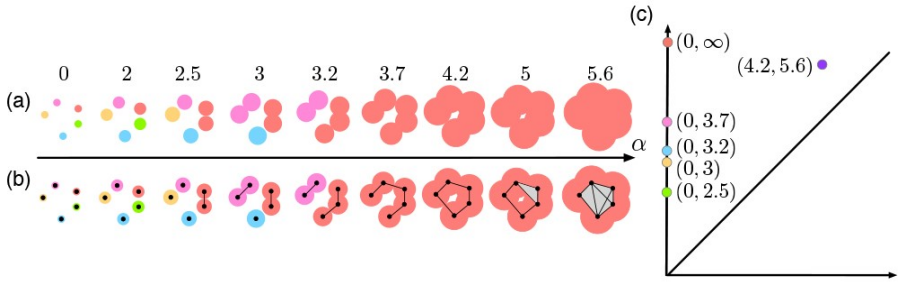

Figura 3: Construcción de un diagrama de persistencia. Se visualiza la nube de puntos a la izquierda y los valores de ε que hacen nacer o morir a alguna componente. En (a) vemos las distintas bolas de radio ε, y en (b) se aprecia superpuesto el grafo que resulta. En (c) pueden verse los puntos correspondientes a las componentes conexas codificados por color. El punto violeta en $(4.2, 5.6)$ corresponde al nacimiento y muerte del loop unidimensional.

3.2.2. Grupos de Homología

En la topología algebraica (que es el área de la matemática que se encarga de estudiar los grupos de homología), la *homología* es una de las formas de "contar agujeros" de espacios topológicos (la otra se llama *homotopía*, y es notablemente más difícil de computar). A la hora de calcular los grupos de homología el trabajo pesado recae sobre el álgebra lineal, pero la brecha de comprensión teórica puede ser difícil de cruzar. La idea de esta pequeña seccion entonces es facilitar esa comprensión utilizando herramientas matemáticas pero sin preocuparse por la formalidad más dura, que se la dejamos a la sección 4.2.

En el contexto de la homología persistente, hay dos direcciones para entender: por un lado, la idea de la persistencia implica varios conjuntos distintos sobre los cuales se quiere computar la homología, y por el otro lado está el cálculo de la homología en sí, para todas las distintas dimensiones de agujeros a considerar. Vamos a olvidarnos por ahora de la persistencia y concentrarnos solamente en formalizar la noción de homología de un solo conjunto.

Imaginémonos entonces que tenemos una nube de puntos, por ejemplo en frente nuestro en 3D, donde algunos puntos están solos, otros puntos están unidos por líneas, algunos tripletes están unidos en un triángulo, y algunos cuartetos en tetrahedros. Por ejemplo en la siguiente imagen (Figura 4):

Este conjunto de puntos, líneas, triángulos, etc, se llama "complejo simplicial", donde un punto es un simplex cero dimensional (0-simplex), una línea es un 1-simplex, un triángulo un 2-

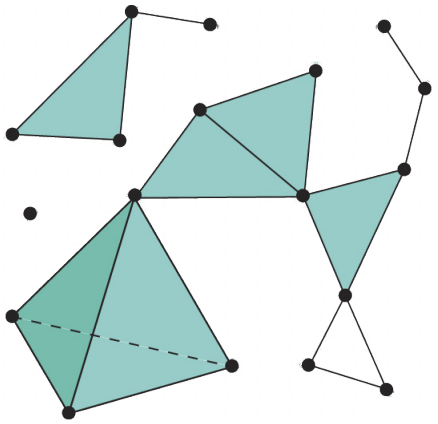

Figura 4: Ejemplo de Complejo Simplicial. Aquí, todas las caras de un simplex son a su vez simples.

simplex, etc. Una característica definitoria de un complejo simplicial es que las caras de los simplex que están presentes también pertenecen al complejo. Entonces en el caso del tetrahedro por ejemplo, no incluimos sólo al tetrahedro en sí sino que también a los cuatro triángulos, seis rectas, y cuatro puntos que lo conforman.

Para poder aprovecharnos de las herramientas del álgebra lineal, necesitamos tener en algún lado un espacio vectorial. Para esto entonces vamos a considerar a todos los simplex que tenemos por cada dimensión, y vamos a nombrarlos σ_i^k; donde para cada dimensión k consideramos a cada simplex indexado por i. En el ejemplo de la imagen tenemos tantos 0-simplex como puntos vemos en la imagen, que serían $\sigma_1^0, \sigma_2^0, ...$, tendríamos todas las líneas (σ_1^1, σ_2^1, ...), ocho triangulos, y un solo tetrahedro.

Recordemos que para un espacio vectorial necesitamos un conjunto base y un conjunto de "números". En este caso, vamos a tener un conjunto por cada dimensión k, y vamos a elegir como conjunto de números al llamado $\mathbb{F}_2$. $\mathbb{F}_2$ es el cuerpo de dos elementos: 0, y 1, donde la suma está definida módulo 2, y la multiplicación está definida como la conocemos: entonces $1+1 = 0$, $1+0 = 1$, $1 \times 0 = 0$, y $1 \times 1 = 1$. Esta elección tiene la ventaja de que los coeficientes pueden interpretarse como la presencia o ausencia de ese simplex en la suma. Vamos a tener objetos del siguiente estilo:

$$0 \times \sigma_1^1 + 1 \times \sigma_2^1 + 1 \times \sigma_3^1 + 0 \times \sigma_4^1 + 0 \times \sigma_5^1 + 1 \times \sigma_6^1$$

En este caso, formamos un nuevo simplex a partir de los simplex originales que teníamos, que es el que resulta de considerar solamente a las líneas indexadas por 2, 3, y 6. Estas sumas de "cosas que no son números" se conocen en general como sumas formales. Si tenemos dos objetos

distintos, por ejemplo $\alpha = \sigma_1^1 + \sigma_3^1$, y $\beta = \sigma_2^1 + \sigma_3^1$, podemos sumarlos juntando los coeficientes de cada 1-simplex y calculando su suma módulo 2:

$$\alpha + \beta = (\sigma_1^1 + \sigma_3^1) + (\sigma_2^1 + \sigma_3^1) = \sigma_1^1 + \sigma_2^1$$

A estos elementos podemos tambien multiplicarlos por elementos de $\mathbb{F}_2$, por lo tanto, el conjunto de sumas formales de los simplex de dimensión k forma un espacio vectorial sobre $\mathbb{F}_2$. A este conjunto lo llamamos C_k.

Tenemos entonces cuatro espacios vectoriales que son no vacíos: C_0, C_1, C_2, y C_3. Para construir los grupos de homología se van a considerar a ciertas funciones entre los espacios vectoriales $\partial_n : C_n \to C_{n-1}$ que van a tomar una suma formal de simplex de n dimensiones y van a enviarlos a la suma formal de sus caras. Por ejemplo, tenemos un solo elemento no nulo presente en C_3: σ^3, el único tetrahedro de la imagen. Este tetrahedro tiene un triángulo en cada cara, llamémoslos σ_1^2, σ_2^2, y σ_3^2. Entonces tendríamos $\partial_3(\sigma^3) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2$, que es la suma formal de sus caras y es un elemento de C_2.

El hecho de que esta función envíe un simplex a sus caras le amerita el nombre "operador de borde". Es importante notar también que volver a tomarle el borde a este nuevo simplex resultaría en el simplex 0: los bordes de los triángulos son segmentos, y cada segmento es eje de dos y solamente dos triángulos, por lo que cuando se sumen vamos a tener a cada segmento repetido dos veces, y estando en $\mathbb{F}_2$, $1 + 1 = 0$, por lo que en la repetición se cancelarían, y la suma sería una suma de ceros. Esto es una de las características centrales de los mapas de borde, y se sintetiza escribiendo $\partial^2 = 0$. Es un hecho general para estos operadores que no depende del simplex al que se le compute el borde ni el conjunto de 'numeros' que se use[7].

Tenemos entonces estos cuatro espacios vectoriales y mapas entre ellos, que conforman una *cadena*:

$$0 \xrightarrow{\partial_4} C_3 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

Para definir el grupo de homología hay dos conjuntos que faltan enunciar: B_k y Z_k. B_k es un conjunto de bordes, y está definido como la imagen del mapa ∂_k. Recordemos que el mapa ∂_k toma k-simplex y los envía a sumas formales de sus caras, entonces en la imagen $\partial_k(C_k)$ están todos los $(k-1)$-simplex que son caras de algún k-simplex (léase: sus bordes). Después, Z_k está definido como el conjunto nulo de ∂_{k-1}; formalmente: $Z_k = \ker(\partial_{k-1})$. Son todos los simplex que ∂_{k-1} envía

[7]Una prueba más general de esto puede encontrarse en el Anexo.

al cero. Hay dos formas de que eso pase: o ya partimos del 0, en cuyo caso $\partial_{k-1}(0) = 0$, o estamos hablando de un elemento de la imagen de ∂_k, o sea $\partial_{k-1}\partial_k(\sigma) = 0$ y estamos en la situación original de ver el borde de un borde.

Tenemos entonces en Z_k a todos los ciclos. Pero estos ciclos pueden ser agujeros o bien pueden ser borde de algún otro simplex. Como solamente nos interesan los agujeros, consideramos el Grupo de cociente Z_k/B_k, esto es: cuento como distintos a todos los ciclos que no son bordes de un simplex de dimensión mayor, que son efectivamente agujeros k-dimensionales presentes en mi complejo simplicial. Ese grupo de cociente es el k-ésimo grupo de homología: H_k.

Este razonamiento se sigue para cualquier dimensión de homología k, y la homología tiene la bella propiedad de que si tenemos un complejo d dimensional, los conjuntos de homología de dimensión mayor a d son siempre 0: no hay agujeros de mayor dimensión que la dimensión total del conjunto.

El salto a la homología persistente es relativamente sencillo. Hasta ahora pensamos en un complejo simplicial fijo sobre el cual desarrollamos los grupos de homología; la extensión viene de la mano de una *filtración*. Pensemos que tenemos una serie de complejos simpliciales, donde cada complejo nuevo va agregando algún k-simplex: le agrega una línea, o un triángulo, o rellena algún espacio que antes estaba solamente rodeado por tres líneas. Esto se conoce como una filtración. Cada uno de los elementos de la filtración es un complejo simplicial en sí, por lo que se le puede calcular la homología, y eso es efectivamente lo que sucede.

En el caso particular de resonancias magnéticas, lo que se tiene es una matriz que resume las distancias entre los nodos de un grafo, y se filtra por esta distancia, incluyendo en cada paso de la filtración a todos los nodos que estén más cerca que la distancia que se está considerando. Después, se "rellenan" todos los simplex posibles entre esos nodos existentes, generando así una familia de complejos simpliciales.

3.3. Vectorización de diagramas

Los diagramas de persistencia son útiles en la medida que pueden resumir invariantes topológicos de varias dimensiones en un mismo gráfico bidimensional, pero esta presentación es inconveniente para incorporarlos en los pipelines estándares del aprendizaje automático. La vectorización de los diagramas de persistencia es una presentación alternativa que resume las mismas

características del diagrama original en un vector. Obtenido el "vector de persistencia" para un diagrama, puede incorporarse a flujos de trabajo estándares tratándolo como un vector de features.

La metodología de vectorización seguida en este trabajo es la prupuesta en PersLay (Carrière et al., 2020). El paper que introduce esto busca incorporar diagramas de persistencia directamente en una arquitectura estándar de red neuronal, para lo que necesitan desarrollar la vectorización de los diagramas. La capa de vectorización la llaman PersLay, y está definida de la siguiente forma:

$$PersLay(D) := \mathbf{op}(\{w(p) \cdot \phi(p)\}_{p \in D})$$

En esta notación, D es un diagrama de persistencia, y p es un punto de ese diagrama. La capa PersLay toma entonces cada punto p del diagrama, y calcula el producto entre $w(p)$ y $\phi(p)$, definidas de la siguiente manera:

- $w(p)$ es una función de peso $w : \mathbb{R}^2 \to \mathbb{R}$ para los puntos del diagrama. El peso asociado a un punto del diagrama se asigna según su cercanía a la diagonal, siguiendo la lógica de que los puntos que representen features van a situarse lejos de la diagonal, mientras que las características que surjan del ruido van a tener una "vida" más corta y se van a situar cerca de la diagonal. Puede ser también una constante $w(p) \equiv 1$.
- $\phi(p)$ es una transformación $\phi : \mathbb{R}^2 \to \mathbb{R}^q$, que mapea a cada punto del diagrama a un vector q-dimensional. Esta función no está predeterminada, pero entre las opciones más comunes y que mencionan en el trabajo original se encuentran la transformación triangular, la gaussiana, y la lineal:
  - Triangular: $\phi_\Lambda : \mathbb{R}^2 \to \mathbb{R}^q$, de forma que $p \mapsto [\Lambda_p(t_1), \Lambda_p(t_2), ..., \Lambda_p(t_q)]^T$ donde $\Lambda_p : t \mapsto$ máx$\{0, y - |t - x|\}$ y los puntos t_i pertenecen a $\mathbb{R}$. Es también conocida como ´´paisaje" o ´´Landscape" en inglés (ver sección 6.3).
  - Gaussiana: $\phi_\Gamma : \mathbb{R}^2 \to \mathbb{R}^q$, $p \mapsto [\Gamma_p(t_1), \Gamma_p(t_2), ..., \Gamma_p(t_q)]^T$, con $\Gamma_p : t \mapsto \exp(-\|p - t\|_2^2 / 2\sigma^2)$ con $\sigma > 0$, y $t_i \in \mathbb{R}^2$, dando resultado a las ´´imágenes de persistencia" (más detalles en la sección 6.3).
  - Lineal: $\phi_L : \mathbb{R}^2 \to \mathbb{R}^q$, $p \mapsto [L_{\Delta_1}(t_1), L_{\Delta_2}(t_2), ..., L_{\Delta_q}(t_q)]^T$, con $L_\Delta : p \mapsto \langle p, e_\Delta \rangle + b_\Delta$, donde e_Δ es un vector dirección y b_Δ un vector de sesgo, y $\Delta_1, \ldots, \Delta_q$ son son líneas en el plano.

Este conjunto de vectores, uno por cada punto del diagrama, se pasa por una operación **op**() es una función invariante ante permutaciones, como el máximo entre los elementos, suma, el k-ésimo valor máximo, etc., también a definir.

3.4. Curvas de Betti

Más allá del diagrama de persistencia, las curvas de Betti son otra herramienta resumen que se desprenden de los grupos de homología. Éstas curvas también son útiles en la tarea de clasificación. Se computan asociando a cada valor de ε de la filtración el número de Betti correspondiente al grupo de homología X_ε, uno por cada dimensión de interés.

Es posible computar el número de Betti directamente desde un diagrama de persistencia. Para el diagrama de dimensión k (D_k) asociado a la filtración X_ε, resulta que:

$$\beta_k(X_\varepsilon) = \#\left\{ (b,d) \in D_k \;\middle|\; b \leq \varepsilon < d \right\}$$

Aquí el símbolo # denota la cantidad de elementos de un conjunto. Nota: Esta definición se desprende de la definición presentada en la Sección 4.2 de los números de Betti, por la cual $\beta_p^i :=$ rango$\{f_p^{i-1,i}(H_p^{\varepsilon_i})\}$. Fijada una dimensión p, un diagrama de persistencia captura los elementos 'vivos' para distintos niveles de ε. Para un ε particular, los elementos que cumplen $b \leq \varepsilon < d$ son aquellos que nacieron antes de ese momento y no murieron todavía. Justamente, el rango de la imagen de $f_p^{i-1,i}(H_p^{\varepsilon_i})$ consiste de aquellos elementos independientes que se encuentran en esa imagen, que el diagrama de persistencia captura en forma del par (b,d).

3.5. Entropía

Los diagramas de persistencia también se han nutrido con ideas de la teoría de la información, importándose el concepto de *entropía*. Partiendo de un diagrama de persistencia $D = \left\{ (b_i,d_i) \;\middle|\; i = 0,\ldots,N \right\}$, utilizando el conjunto de "lifetimes" $L := \left\{ \ell_i = d_i - b_i \;\middle|\; (b_i,d_i) \in D \right\}$ se define la entropía persistente como:

$$H_L = \sum_{i=0}^{N} \frac{\ell_i}{S_L} \log\left(\frac{\ell_i}{S_L}\right) \quad \text{con} \quad S_L = \sum_{i=0}^{N} \ell_i$$

Para cada dimensión de homología entonces puede calcularse la entropía del diagrama de persistencia resultante. En nuestro caso, al trabajar con las dimensiones 0 y 1, podemos asociar a cada diagrama un par de números, correspondiente a las entropías de cada dimensión. También es posible utilizar la entropía como herramienta para filtrar puntos del diagrama de persistencia. Así, puede distinguirse una característica topológica de posible ruido. Más detalle de esto puede encontrarse en Atienza et al. (2016).

3.6. Estabilidad de diagramas de persistencia

Los diagramas de persistencia son estables. Esto quiere decir que frente a pequeñas perturbaciones de la función generadora de los diagramas, los diagramas resultantes también varían poco. Esto es importante porque apunta a que los diagramas de persistencia son robustos ante el ruido: las características topológicas relevadas por los diagramas no sufren demasiado ante perturbaciones en los datos, mientras éstas sean relativamente pequeñas.

En un enfoque más formal, sea $\mathbb{X}$ un espacio topológico, y $f, g : \mathbb{X} \to \mathbb{R}$, y denotemos como $D(f)$ y $D(g)$ a los diagramas de persistencia generados respectivamente con los conjuntos de subnivel de ambas funciones. También, denotamos como $d_B(,)$ a la función de distancia *bottleneck* entre diagramas de persistencia, que es la distancia mínima que puede haber entre los dos puntos más alejados de dos diagramas considerando todas las biyecciones posibles entre ellos. Siguiendo a Cohen-Steiner et al. (2006), tenemos que:

$$d_B\left(D(f), D(g)\right) \leq \|f - g\|_\infty$$

No solamente varían poco los diagramas asociados sino que la distancia entre ambos está acotada por la distancia en L_∞ de las funciones correspondientes. Los resultados de estabilidad de este estilo son útiles porque el relevamiento de datos generalmente introduce ruido; entonces, siempre y cuando el ruido no sea demasiado, podemos acotar el error de aproximación de una función f por su medición ruidosa. Pordemos visualizar esto en la siguiente figura 5.

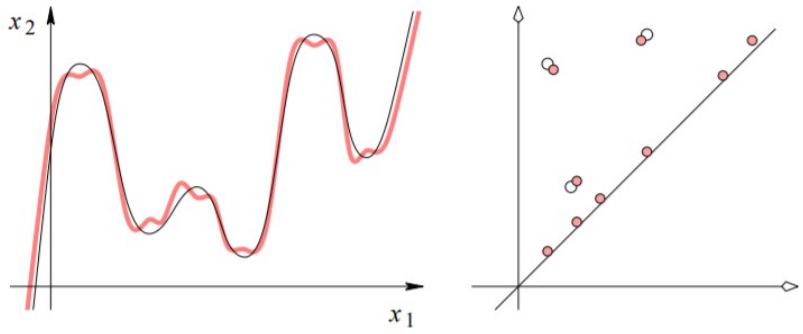

Figura 5: Visualización de dos funciones (izquierda) y sus correspondientes diagramas de persistencia (derecha). Vemos que al ser pequeña la distancia L_∞ entre las funciones, sus diagramas difieren poco también (Edelsbrunner and Harer, 2008, 19)

.

4. Definiciones Matemáticas

En la presente sección se recorren con más formalidad las preliminares matemáticas que son el fundamento de la teoría de la homología, finalizando con la presentación de la homología simplicial en su nivel de formalismo estándar. El contenido de este apartado no es central a los aportes de este trabajo; se incluye en un afán de completitud, y quienes tengan interés están invitades a leerla.

4.1. Preliminares

4.1.1. Topología

Una topología es un primer paso que puede darse para darle estructura a un conjunto. Principalmente esta estructura extra sirve para hablar de **continuidad**. Partimos entonces de un conjunto cualquiera, que llamamos X; una topología τ sobre X es un conjunto que cumple con las siguientes condiciones:

1. El conjunto vacío y X pertenecen ambos a τ
2. Cualquier unión de elementos de τ pertenece a su vez a τ
3. Intersecciones finitas de elementos de τ pertenecen a τ

A los conjuntos que pertenecen a τ se los llama "conjuntos abiertos", porque son una generalización de los conjuntos abiertos estándares en $\mathbb{R}$. La idea es que al tener dos abiertos, por ejemplo $(-1, 1)$ y $(0, 3)$ en los reales, su unión sigue siendo abierta, y su intersección también. De estas propiedades se parte para crear estas tres reglas que definen a una topología, y los conjuntos que estén en esa topología entonces van a comportarse como se comportan los conjuntos que comunmente llamamos abiertos. Para cualquier conjunto, la topología más pequeña que puede construirse es $\{ \emptyset, X \}$, y la mayor es su conjunto de partes $\mathcal{P}(X)$. Por ejemplo, una topología sobre $X = \{ 1, 2, 3 \}$ podría ser $\tau = \{ \emptyset, \{ 1 \}, \{ 2, 3 \}, X \}$.

En una primera aproximación a la noción formal de continuidad generalmente se trabaja sobre $\mathbb{R}$, que es un conjunto con muchas propiedades que facilitan las definiciones. En particular, en $\mathbb{R}$ puede hablarse de la distancia entre dos puntos usando el valor absoluto de su resta. Para definir continuidad de $f : \mathbb{R} \to \mathbb{R}$ en un punto $a \in \mathbb{R}$ entonces, se requiere que para todo $\varepsilon > 0$ que se quiera, exista un $\delta > 0$ que haga verdadera la siguiente implicación: $(|x - a| < \delta \implies |f(x) - f(a)| < \varepsilon)$. Justamente, lo que dice esta definición es que para todos los puntos en una vecindad de $f(a)$ puedo encontrar valores en el dominio que terminen dentro de esa vecindad: si me quedo "cerca" de $f(a)$, me quedo "cerca" de a.

Estas nociones de "cerca" dependen de que sepamos decir qué está cerca y qué está lejos de a y de $f(a)$. Si se tiene una forma de medir distancias, esta noción es clara. Lo que permite una topología es prescindir de esa necesidad y reemplazarla por la noción más general de conjuntos abiertos. Un conjunto X junto con una topología τ conforman un espacio topológico, denotado (X, τ). Ahora, si tenemos dos espacios topológicos (X, τ_X), e (Y, τ_Y), y una función $f : X \to Y$, decimos que f es continua si $f^{-1}(U) \in \tau_X$ para todo $U \in \tau_Y$; recuperando la definición de continuidad de un mapa en un contexto más general. Queda al alcance entonces hablar de continuidad de funciones entre espacios que no tienen la misma estructura que $\mathbb{R}$, e incluso el estudio del conjunto $\mathbb{R}$ equipado con otra topología que no sea la estándar.

4.1.2. Variedades

Una variedad de dimensión d es un espacio topológico $\mathbb{X} = (X, \tau)$ que "localmente se parece" a $\mathbb{R}^d$. Esto es, para cada punto $x \in \mathbb{X}$, puedo acercarme lo suficiente a ese punto como para que se parezca a $\mathbb{R}^d$. La topología acá se usa para formalizar la intuición de "acercarse lo suficiente": deberíamos tener un conjunto abierto A que incluye a x que se parezca a $\mathbb{R}^d$.

El concepto que las variedades engloban no nos es ajeno: en nuestro día a día usamos la idea de que la tierra parece plana si estamos suficientemente cerca al piso, aunque sabemos que la tierra es -a grandes razgos- una esfera. Estando cerca del piso nos movemos en un plano que no está perceptiblemente curvado. El plano ese es $\mathbb{R}^2$, por lo que la esfera sería un espacio topológico de dimensión 2: si bien es una esfera, para cada punto de su superficie te podés acercar lo suficiente como para que se parezca a un plano en dos dimensiones.

Las variedades resultan útiles porque localmente al menos son parecidas a algo que conocemos muy bien, que es $\mathbb{R}^d$, permitiendo entender espacios que en su totalidad son mucho más complejos. En particular, esta propiedad permite que se haga cálculo sobre las variedades, recuperando a las nociones de derivada e integral en un contexto más general, y pudiendo aplicar todas las técnicas que con ellas están asociadas al estudio de las variedades.

Muchas veces se incurre también en la "hipótesis de la variedad", un supuesto del análisis de datos que postula que las variables relevadas provienen de una variedad. Por ejemplo, cuando se ajusta una regresión lineal se asume que los datos provienen de un hiperplano definido por las variables independientes, con un componente de ruido añadido. Ese hiperplano es una variedad.

Cuando se calcula la homología de un conjunto lo que se quiere es capturar su estructura topológica y, cuando el cálculo se hace sobre muestras, se utiliza este supuesto para formalizar la idea de recuperar la homología de la variedad original por medio de muestras u observaciones obtenidas de esa variedad.

4.1.3. Espacio Vectorial

Un espacio vectorial $(V, +, \cdot)$ sobre un cuerpo K consiste de un conjunto V cuyos elementos se llaman *vectores*, una operación llamada suma y otra multiplicación, y se cumple que los vectores pueden ser sumados entre sí y multiplicados por elementos de K, más algunos otros axiomas como conmutatividad y distribución de la multiplicación sobre la suma.

El concepto de espacio vectorial es una generalización de los vectores que normalmente se conocen en las clases de matemática: se habla de los vectores $v = (1, 2)$ y $u = (3, 0)$, por ejemplo, y se calculan expresiones como:

$$3v - u = 3\begin{pmatrix}1\\2\end{pmatrix} - \begin{pmatrix}3\\0\end{pmatrix} = \begin{pmatrix}3\\6\end{pmatrix} - \begin{pmatrix}3\\0\end{pmatrix} = \begin{pmatrix}3-3\\6-0\end{pmatrix} = \begin{pmatrix}0\\6\end{pmatrix}$$

Las ideas son las mismas, solo que en vez de hablar de vectores como "columnas de números", se piensan como elementos abstractos v, y en vez de multiplicarlos por números reales, como el 3, se multiplican por "números" de un conjunto[8] K. Los espacios vectoriales nos van a servir como base para construir los grupos de homología. Y partir de espacios vectoriales va a permitir utilizar las herramientas del álgebra lineal para computar las homologías a distintos valores de ε.

4.1.4. Grupos

Para introducir el concepto de Grupo partamos por ejemplo de un triángulo equilátero. Una de las cosas que puede hacerse con ese triángulo es reflejarlo sobre su eje vertical medio. Si se lo refleja así tenemos nuevamente un triángulo. También se lo puede rotar un tercio de vuelta a la derecha para obtener una versión rotada del mismo triángulo. Pueden apreciarse en la Figura 6.[9] Estas acciones pueden combinarse: rotar dos veces, rotar y reflejar, reflejar y rotar, etc. Notemos que no siempre da lo mismo el orden: rotar y reflejar por el eje vertical no es lo mismo que reflejar primero y rotar después; y que algunas transformaciones son lo mismo que no hacer nada: reflejar dos veces y rotar tres veces vuelven el triángulo a su posición original.

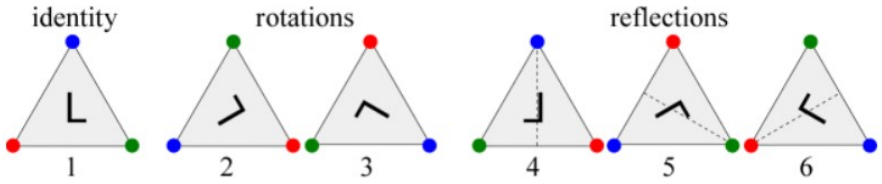

Figura 6: Ejemplo de las simetrías de un triángulo equilátero

Pensemos en el conjunto de todas las simetrías posibles de un triángulo equilátero, y llamemos çomposición.ª la acción de hacer una transformación después de la otra, combinando dos simetrías distintas. Estas simetrías junto con la forma de combinarlas forman un **grupo**.

Más formalmente, un grupo $G = (A, \circ)$ es un conjunto A junto con una operación $\circ$ en elementos de A que cumple:

1. $a \circ (b \circ c) = (a \circ b) \circ c$ para cualesquiera elementos $a, b, c \in G$

[8] Técnicamente no es cualquier conjunto, sino que son conjuntos específicos que se llaman Cuerpos. Lo más fácil es pensarlo como $K = \mathbb{R}$ directamente. Puede relajarse el requerimiento de que K sea un cuerpo a requerir que sea solamente un *anillo*, en cuyo caso a la estructura resultante en vez de llamarla Espacio Vectorial se lo llama K-módulo

[9] En *Triangle Fractals - Agnes Scott College*. URL: `http://larryriddle.agnesscott.org/ifs/siertri/triangleVariation.htm`

2. Existe un elemento $e \in G$ que cumple que $e \circ a = a \circ e = a$ para todo $a \in G$, llamado elemento identidad.

3. Para todo elemento $a \in G$ existe un $b \in G$ tal que $b \circ a = a \circ b = e$, denotado como $b = a^{-1}$, llamado inversa de a.

Los grupos en particular que nos van a interesar cumplen una propiedad extra: $a \circ b = b \circ a$ para todos los $a, b \in G$. Los grupos que cumplen esta propiedad se llaman grupos *abelianos* o conmutativos. El caso del triángulo es un ejemplo de un grupo no-abeliano. El arquetípico ejemplo de un grupo abeliano es el grupo de los enteros con la suma: $(\mathbb{Z}, +)$. El elemento identidad es el 0, y la inversa de un entero es el negativo de ese entero: la inversa del 4 es -4 y vice versa.

Otro ejemplo de grupo abeliano es el conjunto de las matrices reales de $n \times n$ y la suma. En este caso la suma es conmutativa porque al hacerse componente a componente, hereda la conmutatividad de la suma entre números reales en sus entradas; y el elemento identidad es la matriz que tiene todos ceros. En cambio, las matrices de $n \times n$ de números reales bajo la multiplicación no forman un grupo: no todas las matrices tienen inversa multiplicativa. Hay que restringirse a las matrices con determinante no nulo para tener un grupo bajo la multiplicación. Un "subgrupo" de estas matrices son aquellas de determinante 1, que forman un grupo en sí mismas, conocido como $\mathrm{SL}_n(\mathbb{R})$: el grupo especial lineal de matrices de $n \times n$ sobre $\mathbb{R}$. Este grupo es especial porque define las rotaciones y reflecciones de $\mathbb{R}^n$ en sí mismo que no distorsionan ángulos o distancias entre puntos.

4.1.5. Grupos de cociente

Los grupos de cociente son también grupos, pero que se caracterizan por cómo son construídos. En este caso tomemos como ejemplo al reloj de 12 horas. Si salimos en un viaje de tres horas a las 11, la hora de llegada son las 2. ¿Por qué la hora de llegada no son las 14? Esta idea de contar hasta 11 y voler al 0 es lo que se conoce como "suma módulo 12", y es la suma que se requiere para que el conjunto $\{0, 1, ..., 11\}$ sea un grupo. En este caso la identidad es el 0, y, por ejemplo, la inversa del 1 es el 11, y la del 7 es el 5, porque $11 + 1 = 12$, y $7 + 5 = 12$; y el 12 se identifica con el 0.

Para construirlo partimos del grupo de los números enteros con la suma: $(\mathbb{Z}, +)$. Pensemos en el conjunto $12\mathbb{Z}$ de múltiplos de 12: $\{\ldots, -12, 0, 12, 24, 36, \ldots\}$. La clave consiste en pensar a un número entero cualquiera como un múltiplo de 12 más un resto, y concentrarse en ese resto.

Por ejemplo, 27 es $2 \times 12 + 3$. ¿Cuál es el resto de dividir 27 en 12? Es 3. Esto se escribe $27 \equiv 3$ mód 12. Pensemos que al dividir por 12, los únicos restos posibles son los enteros desde el 0 hasta el 11. Ese conjunto de restos, equipado con la suma módulo 12 es justamente el grupo cociente de $\mathbb{Z}$ y $12\mathbb{Z}$, y se escribe $\mathbb{Z}/12\mathbb{Z}$.

Esta construcción con los enteros funciona para cualquier número entero n. Como idea general, podemos pensar en el grupo de cociente G/H: en este caso "miramos" a los elementos de G luego de "quitarles su parte de H", y se consideran equivalentes dos elementos de G si difieren por un elemento de H.

Otro ejemplo posible es el grupo de los reales $(\mathbb{R}, +)$ con su subgrupo $(\mathbb{Z}, +)$: el grupo $\mathbb{R}/\mathbb{Z}$ es equivalente a un círculo. A cada número real se lo piensa como la suma de su parte entera k más un número $r \in [0, 1)$, que sería su parte decimal, y se considera solamente la parte decimal. Más aún, el número real 1 resulta quedar igual al 0 (así como las 12hs es igual que las 0hs), entonces si se "camina" por el intervalo $[0, 1]$, cuando se cruza el 1 se reaparece por el lado del 0, que es lo mismo que pasa en un círculo.

4.2. Homología simplicial

Para calcular la homología persistente de una nube de puntos es necesario poder calcular los grupos de homología a todas las escalas posibles. Para lograr esto, no se trabaja con los puntos aislados, sino que se trabaja con un complejo simplicial asociado a esa escala.

Un simplex de dimensión k en $\mathbb{R}^n$ es el cubrimiento convexo de $(k + 1)$ puntos linealmente independientes en $\mathbb{R}^n$, $(n > k)$. Es decir, dados vectores $\{v_0, v_1, ..., v_k\}$ de $\mathbb{R}^n$, el k-simplex asociado sería $\left\{ \sum_{i=0}^{k} t_i v_i \;\middle|\; t_0 + t_1 + ... + t_k = 1,\ t_i \geq 0 \right\}$

Los simplex que nos van a interesar son los simplex estándares, que son aquellos que se componen de los vectores base estándares de $\mathbb{R}^n$. El k-simplex estándar, Δ_k, se define entonces como:

$$\Delta_k = \left\{ \sum_{i=0}^{k} t_i e_i \;\middle|\; t_0 + t_1 + ... + t_k = 1,\ t_i \geq 0, (e_i)_j = \delta_{ij} \right\}$$

Entonces, un 0-simplex es un punto, un 1-simplex es una línea, un 2-simplex un triángulo, un 3-simplex un tetrahedro, y así sucesivamente. Dado un conjunto finito X, un complejo simplicial $\mathcal{K}$ sobre X es un conjunto de subconjuntos de X tal que:

1. Si $\sigma \in \mathcal{K}$ y $\pi \subset \sigma$, entonces $\pi \in \mathcal{K}$
2. Si $\sigma, \pi \in \mathcal{K}$, entonces $\sigma \cap \pi$ es vacía o es una cara de ambos.

Básicamente lo que esta definición dice es que todos los simplex que se encuentran en $\mathcal{K}$ están formados por simples de dimensión menor (que también están en $\mathcal{K}$). Definimos entonces el conjunto de k-cadenas de $\mathcal{K}$ sobre X como:

$$C_k(X) = \left\{ \sum_i a_i \sigma_i^k \;\middle|\; \sigma_\alpha^k : \Delta_k \to X,\ a_i \in \mathbb{F}_2 \right\}$$

Los elementos de $C_k(X)$ se llaman k-cadenas. Una k-cadena es una suma formal $\sum a_i \sigma_i^k$ donde los σ_i^k son simplex y los a_i son coeficientes, que para este trabajo se tomaron como $a_i \in \mathbb{F}_2$, el cuerpo de dos elementos. Esta elección de cuerpo es estándar en la topología computacional [10]. La diferencia es que al elegir un cuerpo, $C_k(X)$ es un espacio vectorial, y si se elige un anillo conforma un módulo.

Se definen mapas $\partial_n : C_n(X) \to C_{n-1}(X)$ por su acción sobre los generadores σ_α^k, con $\partial_n(\sigma_\alpha^k) = \sum_{i=0}^n \sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]}$, donde la notación $\sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]}$ refiere al simplex que resulta de removerle el vértice v_i a σ_α. La acción de estos mapas sobre una k-cadena cualquiera se obtiene expresando cada cadena en términos de sus generadores y usando la linealidad del mapa sobre la suma.

Estos mapas cumplen que $\partial_{n-1} \circ \partial_n = 0$ siempre, por lo que $Im(\partial_n) \subseteq Ker(\partial_{n-1})$. Se definen entonces los conjuntos $Z_n = Ker(\partial_{n-1})$, y $B_n = Im(\partial_n)$, llamados respectivamente "ciclos" y "bordes". Lo que nos interesa en la homología es contar todos los ciclos que no resultan bordes de un simple de dimensión mayor, porque sino no serían agujeros, entonces se define al n-ésimo grupo de homología como: $H_n(X) := Z_n/B_n$. Es el grupo de cociente de los ciclos descontando los bordes.

Recordemos que para este trabajo para cada sujeto partimos de un conectoma que es representado como un grafo no direccionado y contamos con la distancia entre los distintos nodos en

[10] Podría elegirse otro cuerpo como $\mathbb{Q}$, $\mathbb{R}$, o cualquier cuerpo finito $\mathbb{F}_p$ ara p primo. También se puede generar un grupo abeliano libre sobre los generadores que sería como elegir $\mathbb{Z}$, o elegir cualquier anillo unitario R

forma de una matriz. Nos va a interesar entonces calcular filtraciones de estos grafos para luego calcular sus grupos de homología, entre otras cosas. Hay dos formas principales de pasar de una nube de puntos a un complejo simplicial: usando complejos de Čech y usando complejos de Vietoris-Rips.

Dado un grafo $G = (V, E)$ y una función de peso $w : V \times V \to \mathbb{R}$, definimos la filtración de Vietoris-Rips de G como la familia de subgrafos indexados por ε: $G_\varepsilon = (V_\varepsilon, E_\varepsilon)$. Aquí, $V_\varepsilon = V$ para todo ε, y E_ε contiene todos los ejes cuyo largo sea menor o igual a ε. El complejo $Cl(G_\varepsilon)$ es el complejo simplicial asociado al grafo filtrado, donde para cada k se incluyen todos los k-grafos completos y cada cara es a su vez un $(k-1)$-simplex. Esto es: se incluyen todas las k-cadenas posibles donde cada simple $\sigma = \{\sigma^k_{i_1}, \ldots, \sigma^k_{i_m}\} \in Cl(G_\varepsilon)$ cumple que $d(\sigma^k_{i_j}, \sigma^k_{i_l}) < \varepsilon$ para todos los i_j e i_l. Una forma de pensar esto es utilizando la función de pesos w, y considerando a los conjuntos de subnivel, definiendo $E_\varepsilon := w^{-1}((-\infty, \varepsilon])$, las preimágenes de las semirectas $(-\infty, \varepsilon]$. Por fuera del contexto particular de grafos puede tomarse cualquier función $f : \mathbb{X} \to \mathbb{R}$ y considerar las preimágenes de la semirectas como las filtraciones sobre las cuales calcular homologías. En cualquier caso, vale notar que si $\varepsilon_1 < \varepsilon_2$, entonces $f^{-1}((-\infty, \varepsilon_1]) \subset f^{-1}((-\infty, \varepsilon_2])$.

La filtración de Čech es similar, con la diferencia que considera en cada nivel de filtración ε solamente a los simplex donde todos los vértices estén a distancia menor o igual a ε. Esto hace que en algunos casos, para el mismo nivel de filtración, el complejo de Vietoris-Rips incluya algunos simplex que el complejo de Čech no. En los algoritmos implementados por las librerías que se utilizan en este trabajo se utiliza la filtración Vietoris-Rips.

Independientemente de la filtración que se elija entre estas dos, en las aplicaciones que nos interesan hay solamente finitos epsilon $\varepsilon_1 < \varepsilon_2 < \ldots < \varepsilon_n$ que generan un cambio en $Cl(G_\varepsilon)$: aquellos donde se agrega algún k-simplex al complejo simplicial. Tenemos entonces una filtración finita de complejos simpliciales

$$Cl(G_{\varepsilon_1}) \xhookrightarrow{i_{1,2}} Cl(G_{\varepsilon_2}) \xhookrightarrow{i_{2,3}} \ldots \xhookrightarrow{i_{n-1,n}} Cl(G_{\varepsilon_n})$$

donde los mapas $i_{m,m+1}$ denotan inclusión.

Para cada conjunto $Cl(G_{\varepsilon_i})$ y cada dimensión de homología p, tenemos el grupo de homología asociado $H_p(Cl(G_{\varepsilon_i})) =: H_p^{\varepsilon_i}$. A su vez, la inclusión de la filtración provee homomorfismos $f_p^{i-1,i} : H_p^{\varepsilon_{i-1}} \to H_p^{\varepsilon_i}$. Los números de Betti comentados anteriormente están definidos como el rango de las imágenes de estos homomorfismos: $\beta_p^i := \text{rango}\{f_p^{i-1,i}(H_p^{\varepsilon_i})\}$.

5. Dataset

5.1. Dataset IMPAC

Se cuenta con datos de fMRI de 1150 sujetos, que forman parte del dataset público del IMPAC[11], y un conjunto de features estructurales extraídas y guardadas en un archivo CSV, donde cada fila le correponde a un sujeto y cada columna a un feature específico. Dentro de la competencia, el conjunto de testeo constó de 1003 casos adicionales a los 1150 originales que luego de la competencia no se disponibilizaron públicamente.

Para cada sujeto también se cuenta con las señales de fMRI correspondientes a distintas parcelaciones cerebrales. Las parcelaciones disponibles son:

- Parcelaciones BASC con 64, 122, y 197 regiones;
- Parcelación Craddock Scorr Mean;
- Parcelación anatómica Harvard-Oxford;
- Atlas funcional MSDL;
- Atlas Power 2011;

Todas estas son estándares en la literatura (Ver (Bellec et al., 2010; Craddock et al., 2012; Varoquaux et al., 2011; Power et al., 2011)). Las series de tiempo obtenidas en el fMRI se encuentran preprocesadas, y las series temporales cerebrales divididas según las Regiones de Interés están disponibles. Las series de fMRI por parcelación constan de 120 observaciones secuenciales en el tiempo, con tantas variables como secciones tenga la parcelación utilizada. En la figura 7 podemos ver ejemplos de dos sujetos, uno con diagnóstico positivo otro negativo, con los valores promedios y desvíos de las medidas de actividad en cada momento del tiempo.

Las distribuciones de los casos entre estos 1150 sujetos por la variable de interés, por sexo, y por edad, están disponibles públicamente en la página del IMPAC [12] y se muestran en la Fig. 8. Puede apreciarse también la distribución del conjunto privado de testeo.

[11] `https://paris-saclay-cds.github.io/autism_challenge/`
[12] `https://paris-saclay-cds.github.io/autism_challenge/`

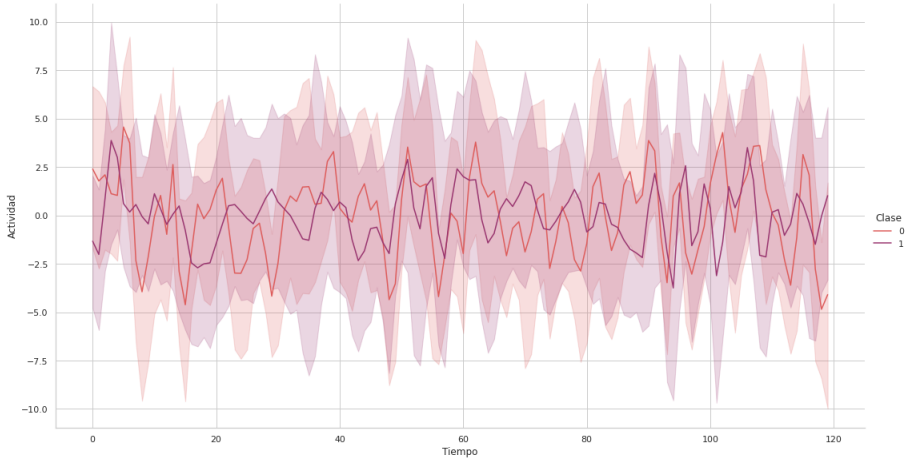

Figura 7: Ejemplos de fMRI del conjunto de entrenamiento por clase. Se observa el promedio de todas las series con una banda de un desvío estándar. La clase negativa está en rojo y la positiva en violeta.

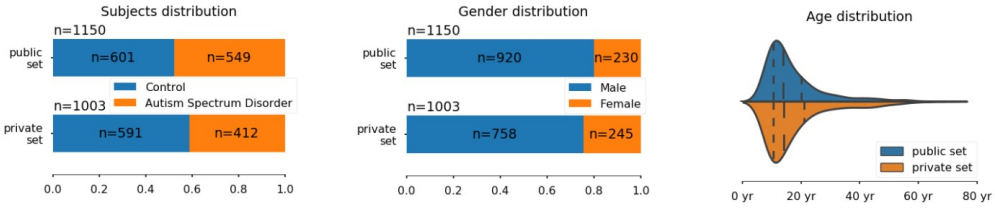

Figura 8: Distribución de datos del IMPAC según la presencia de Autismo, sexo, y edad, de los conjuntos de entrenamiento (público) y testeo (privado).

Las variables que resumen medidas anatómicas son 207. Entre ellas pueden encontrarse el área lateral occipital de los hemisferios derecho e izquierdo, el área supramarginal, los volúmenes de la corteza izquierda y corteza derecha, el volumen estimado total intracranial (eTIV), volumen de materia gris, el Volumen cerebral normalizado computado utilizando la segmentación subcortical de FreeSurfer (suite de procesamiento de MRIs), y el grosor de la corteza y áreas de los hemisferios derecho e izquierdo, también procesado con el software FreeSurfer. Estas variables concentran el 99,9 % de su variabilidad en las primeras cinco componentes principales de la descomposición PCA del conjunto de entrenamiento. Puede apreciarse el gráfico del porcentaje de variabildiad explicada por componente principal en el gráfico 9.

Se presenta un resumen de las características expuestas hasta ahora en la tabla 1, y puede verse a continuación una proyección de UMAP de las variables anatómicas coloreadas por clase en el gráfico 10.

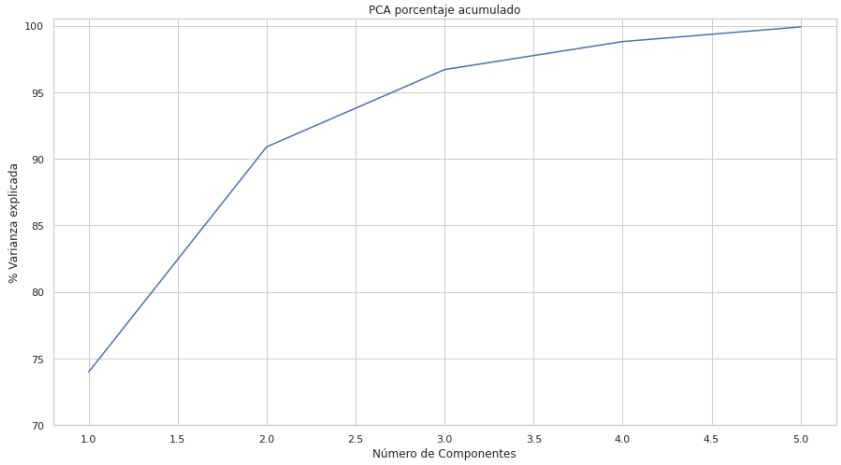

Figura 9: Variabilidad explicada acumulada por componente principal de las features anatómicas de los sujetos en el conjunto de entrenamiento. Se observa que con los primeros tres componentes se captura más del 95 % de variabilidad.

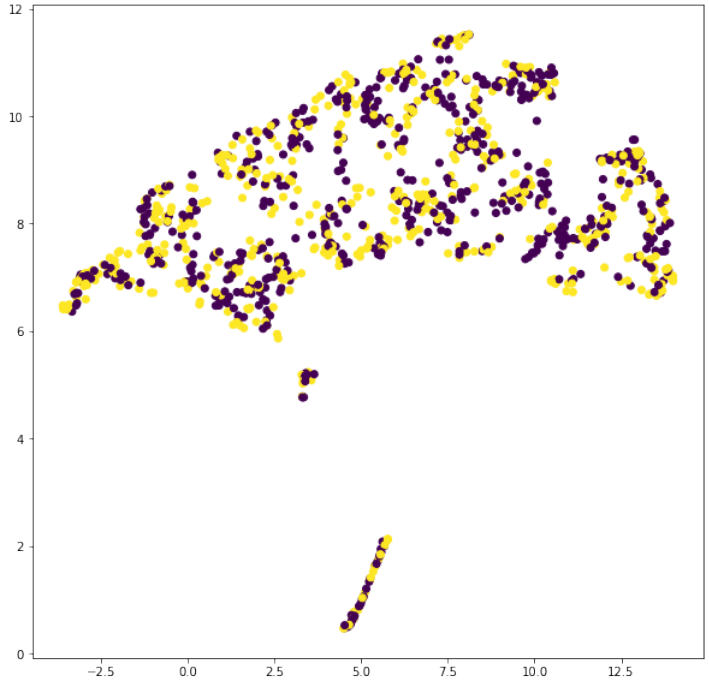

Figura 10: Proyección UMAP de features anatómicas según diagnóstico. La clase positiva es amarilla mientras que la negativa es violeta. No se visualiza una separación clara de clases.

5.2. Datos ABIDE comparativos

Para poder complementar el análisis sobre el dataset IMPAC se han replicado algunos pipelines con otro gran dataset de resonancias magnéticas funcionales en reposo de Autismo,

Aspectos	**Dataset IMPAC**
Número de casos	1150
Número de variables anatómicas	207
Parcelaciones	7
Observaciones por fMRI	120

Cuadro 1: Resumen Dataset IMPAC

que es el dataset ABIDE. En este caso, se utilizó el ABIDE Preprocessed, que contiene disponibles resonancias preprocesadas y etiquetadas de sujetos con y sin TEA. A su vez, utilizando la API disponible de consulta, se buscó concentrar el foco en un subconjunto de menor variabilidad de la que presenta IMPAC. Se optó entonces por buscar aquellas instancias disponibles de sujetos que tuvieran entre 18 y 20 años. Se cuenta entonces con 40 resonancias magnéticas funcionales adicionales, mitad y mitad pertenencientes a la clase positiva y la clase control. Para más información de los preprocesamientos realizados a los datos crudos, referirse a `http://preprocessed-connectomes-project.org/abide/index.html`

6. Métodos

En esta sección se detallan los métodos que se siguieron en la conducción de experimentos. En líneas generales el proceso toma la resonancia magnética funcional de cada sujeto y calcula la similaridad de actividad entre sus regiones cerebrales. Es necesario luego convertir esta matriz de similaridad a una matriz de distancias para poder extraer las características topológicas explicadas en la sección 3. Esta información puede luego resumirse en un vector para utilizarse como entrada de los modelos predictivos.

6.1. Conformación del conectoma

Partimos de una resonancia magnética funcional por cada sujeto. Esto se guarda como un conjunto de p series temporales, donde cada serie representa la actividad medida en una sección del cerebro determinadas por la parcelación elegida. La forma en la que se trabaja en general con este tipo de datos es calculando una matriz de similaridad entre las secciones, donde la entrada (i, j)-ésima corresponde a la similaridad entre la sección i y j de la parcelación. En el estado del arte actual a esta matriz se la suele vectorizar, descartando todas las entradas que no estén en el

triángulo superior (sin la diagonal). Este vector puede luego incorporarse al pipeline estándar de modelado, ya sea para una regresión logística o un modelo de Machine Learning.

Aparte de vectorizarla puede resumirse la información en un grafo no direccionado con peso asignándole un nodo a cada sección de la parcelación y pesando los ejes según las entradas de la matriz de similaridad. De esta forma se asocia a cada sujeto con un grafo que resume la conformación de su cerebro, donde la cercanía entre secciones ya no es anatómica sino que depende de los niveles de actividad conjunta.

Hay tres formas de calcular esta matriz de similaridad:

1. Calculando la matriz de correlación entre los nodos. Las series temporales se centran en 0 restándoles su media, y la i, j-ésima entrada de la matriz contiene la correlación de Pearson entre la serie i y la serie j de la parcelación. Esta medida es la más simple de las tres y tiene la ventaja de ser acotada, facilitando la comparación entre las distintas entradas de la matriz. Asumiendo una hipótesis nula de ausencia de correlación puede calcularse la significatividad de cada entrada y así descartar aquellas que no alcancen un nivel de significatividad predeterminado con el fin de reducir el ruido en las conexiones. En las entradas que no resulten significativas se coloca un 0, efectivamente desconectando los nodos correspondientes a ese eje.

2. El segundo método es la correlación parcial. A diferencia de la correlación anterior, la correlación parcial no calcula la asociación entre las series temporales directamente, sino que lo hace sobre los residuos de la estimación de las series con todas las demás. Esto es: a cada serie i, se la regresa contra las otras y se guardan los residuos de esa estimación. Luego, se calcula la correlación entre los residuos de la serie i con los residuos del mismo procedimiento con la serie j. En la práctica, la correlación parcial entre las series i y j se encuentra en el (i, j)-ésimo elemento de la inversa de la matriz de covarianza de todas las series; entonces solo es necesario calcular las covarianzas e invertir la matriz[13], ahorrando el costo computacional del cómputo de las regresiones.

3. Se puede también utilizar la proyección tangencial de las matrices de covarianza. Lo que motiva este médodo es la observación de que el conjunto de matrices de covarianza (y las matrices positivas definidas en general) no forman un espacio euclideano, lo que vuelve difusa la interpretación de las comparaciones entre distintas matrices. Lo que sí ocurre es que este

[13]Puede verse una prueba de esto en https://stats.stackexchange.com/questions/140080/why-does-inversion-of-a-covariance-matrix-yield-partial-correlations-between-ran

conjunto es una variedad diferenciable[14]. Se puede entonces tomar una matriz de referencia M y calcular el espacio tangente a M, que es un espacio vectorial (y euclídeo porque es el producto cartesiano de Reales). Se procede luego a proyectar a todas las matrices de covarianza que tengamos sobre este espacio, así obteniendo matrices que pueden sumarse y restarse de forma cerrada. En general esta matriz de referencia M se toma como la media aritmética entre las matrices de interés. A diferencia de los dos métodos anteriores, entonces, la descomposición tangencial requiere computar la media de todas las matrices involucradas antes de calcular las proyecciones.

6.2. Del conectoma al Diagrama de Persistencia

Para construir el diargama de persistencia se utilizan las preimágenes de la función de pesos del grafo. Si bien la técnica funciona para cualquier función que tenga imagen en los reales y en particular funciona también para las medidas de similaridad (correlación, correlación parcial, y descomposición tangente), como el cálculo de la homología persistente aplicado en este trabajo toma los conjuntos de subnivel $(-\infty, \varepsilon]$, todos los nodos aparecen primero en las similaridades negativas, y se pierden los aportes de las similaridades positivas. Es por eso que antes de calcular el diagrama de persistencia es necesario transformar las matrices de similaridad en matrices de distancia. Para hacer esto durante la fase exploratoria del proyecto se trabajó con dos transformaciones:

1. La distancia coseno: dada una similaridad s, acotada en $[-1, 1]$ la distancia coseno asociada es $d = \sqrt{2(1-s)}$.
2. La distancia dinámica: dada una similaridad s, acotada en $[-1, 1]$, la distancia dinámica es $d = 1 - |s|$.

La aplicación de estas transformaciones a las matrices de correlación y correlación parcial es directa. Con respecto a las matrices tangenciales, se tomó primero el valor absoluto máximo de todas las entradas de las matrices y se dividieron todas las matrices por ese valor, acotándolas al $[-1, 1]$.

[14]El conjunto de todas las matrices simétricas es un espacio vectorial de dimensión finita (sobre algún cuerpo K, asumamos los Reales); y fijando vectores base se puede cubrir con un mapa global, convirtiéndolo en una variedad. En particular, las matrices positivas definidas tienen autovalores estrictamente positivos, por lo que hay vecindades abiertas a estas matrices donde se sigue cumpliendo la condición de ser positivas definidas. Al ser un subconjunto abierto de una variedad, las matrices positivas definidas conforman una variedad en sí mismas.

En términos de el marco teórico de la homología, trabajar con la matriz de distancias es equivalente a generar el complejo simplicial conectando todos los nodos a distancia menor a ε. La diferencia es que aquí no partimos en sí de los puntos en el espacio y una función de distancias de a pares, sino que la matriz contiene directamente la distancia entre los puntos. Así, ver qué nodos están a menos de ε de distancia es equivalente a filtrar la matriz de distancias por aquellas entradas que sean menores que ε.

Existe también una homología persistente que computa al mismo tiempo la homología utilizando los conjuntos de subnivel y los de supranivel, $[\varepsilon, \infty)$. Esta técnica no está implementada todavía, y al momento solo pude encontrar una referencia de su uso (Edelsbrunner and Harer, 2010, 192), y no posee todavía implementaciones en Python, por lo que se eligió el camino ya mencionado.

De cualquier manera, se computaron los diagramas de persistencia de dimensiones 0 y 1 partiendo de las matrices de distancia. Las dimensiones de 2 en adelante no se incluyeron en el pipeline general ya que el tiempo de cómputo es significativamente mayor al punto que imposibilita realizar la cantidad de transformaciones necesarias.

6.3. Vectorización del Diagrama de Persistencia

Durante la exploración se utilizaron principalmente las vectorizaciones de los diagramas de persistencia para poder evaluarlos en conjunto con, por ejemplo, proyecciones de Componentes Principales (en contraste con utilizar directamente los diagramas). La técnica utilizada está detallada en la Sección "TDA y homología persistente - Vectorización de diagramas". Durante la exploración se utilizaron distintos parámetros para evaluar su efecto en las proyecciones de los diagramas. Previo a la transformación, los diagramas se escalaron para acotar los valores al $[0, 1]$ y se añadieron puntos en la diagonal para equiparar la cantidad de puntos por diagrama. Se descartaron también los puntos con alguna coordenada infinita. Aquí se encuentran dos grandes formas de vectorizar los diagramas, uno conocido como la transformación *Landscape*, y otra como *Imágenes de Persistencia*.

La transformación Landscape es más fácilmente entendida con un diagrama. Veamos la Figura 11. Lo que hace esta transformación es "estirar" el diagrama de persistencia para que ocupe todo el primer cuadrante, y a cada punto lo toma como la punta de una "montaña", y luego toma el máximo de todas las montañas para conformar el Landscape o paisaje. En la figura referenciada[15], la línea azul sería el paisaje que resulta, y se vectoriza registrando la altura del paisaje en ciertos

[15]Figura en "Using Topological Data Analysis to Process Time-series Data: A Persistent Homology Way", DOI 10.1088/1742-6596/1550/3/032082

puntos. Para este trabajo se muestrearon 100 puntos consecutivos e igualmente espaciados para la vectorización.

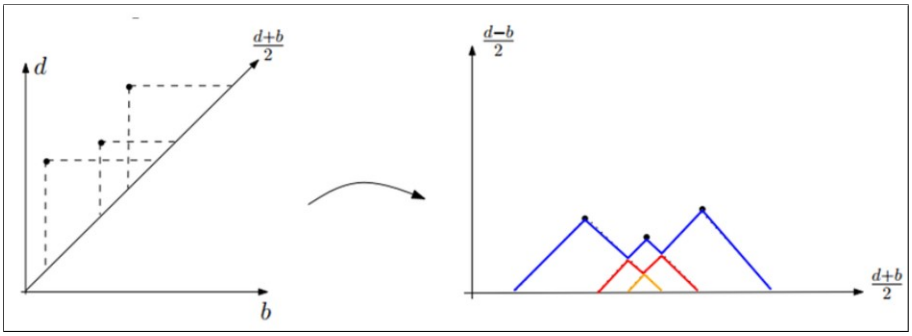

Figura 11: Representación gráfica de la transformación Landscape o paisaje. Primero se transforma al diagrama para que sus puntos ocupen todo el primer cuadrante y luego se toma el máximo de las "montañas" con picos en cada punto.

Las imágenes de persistencia parten de la misma transformación del diagrama para que ocupe todo el primer cuadrante, pero en vez de considerar el paisaje se le superopne a cada punto una Gaussiana, de forma tal que el cuadrante entero puede representarse como una imagen y usar la altura de las funciones en ese punto como nivel de color. Podemos encontrar en **??** una imagen de este proceso, expuesta aquí en la figura 12. En esta transformación se vectorizó directamente la imagen, pero también podría tomarse como input para una Red Neuronal Convolucional.

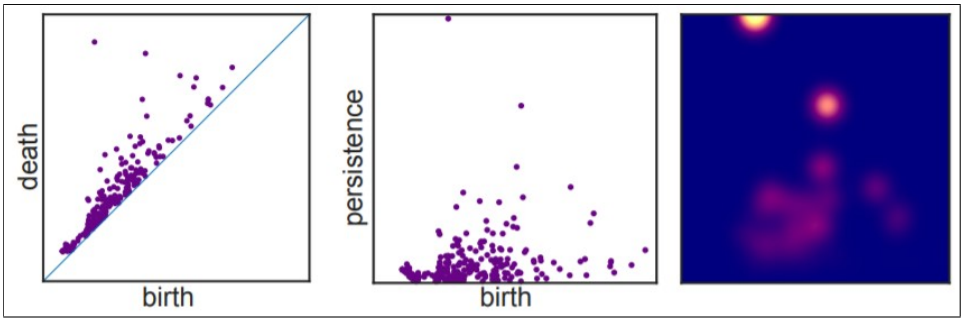

Figura 12: Representación gráfica de una imagen de persistencia obtenida de un Diagrama. La contribución de cada Gaussiana está escalada según su distancia a la diagonal. Partiendo de un diagrama (izq.), se transforma para ocupar todo el cuadrante (medio), y se genera la imagen (der.)

6.4. Prueba de Concepto

En muchos dominios es posible utilizar directamente los datos en la clasificación (descontando algún preprocesamiento o limpieza de los mismos). En este caso, el camino desde el dato crudo de fMRI hasta el vector de features tiene varios pasos intermedios. Todos estos pasos están sujetos a errores y la clasificación puede verse afectada por ellos. Por eso, y con el objetivo de asegurar un correcto funcionamiento del pipeline de preprocesamiento, se generaron observaciones de dos familias diferentes. Para poder replicar el preprocesamiento en su totalidad, se simularon series temporales que tomarían el lugar de las series temporales de actividad medidas en los fMRI. Sobre ellas se aplicaron los mismos pasos que en la fase exploratoria para los datos de resonancia magnética.

Las dos familias se conformaron de la siguiente forma: En la primera, la mitad de los canales se simularon utilizando ruido aleatorio distribuído como una Normal $N(0,1)$. La otra mitad combinó una señal oscilatoria, una ley de potencias, y un ruido normal. Indexando con i las primeras observaciones, se utlizaron los siguientes parámetros.

- Con $h \sim U(\{7,10,15\})$, $\kappa \sim U(\{-0{,}1,-0{,}5,-1,-1{,}5\})$, $\sigma \sim U(\{0{,}5,1,2,3,4,5\})$
- $O(x)$ siendo una señal oscilatoria de frecuencia x
- $P(\lambda)$ siendo una señal de potencias de parámetro λ
- $\Sigma(s)$ siendo una señal de ruido normal con media cero y varianza s
- Se generaron las señales (indexadas por i): $\big((O(h)+P(\kappa))i+\Sigma(\sigma)\big)_i$

En la segunda familia (y utilizando la misma notación que para la primera), la primera mitad de las series indexadas por i se crearon según la ley $(P(\kappa)i+\Sigma(\sigma))$, donde esta vez $\kappa \sim U(\{-1,-2,-3,-4\})$ y $\sigma \sim U(\{0{,}5,1,2,3,4,5\})$. Y para la segunda mitad, se generaron series de ruido blanco normal de media 0 y varianza sampleada uniformemente de $\{0{,}5,1,2,3\}$.

En cada caso se generaron observaciones de 100 series temporales con 30 observaciones cada una, mitad y mitad de la familia 1 y 2. En general, la (1) tiende a tener los puntos de dimensión 1 agrupados en un solo cluster, y la homología 0 se limita a la mitad superior del eje y. En cambio, la familia (2) tiende a tener los puntos de dimensión 1 en dos grupos, uno más concentrado y otro más disperso localizado más cerca del origen, y los puntos de dimensión 0 ocupan en general

toda la línea vertical arriba del cero. Se pueden ver ejemplos a continuación de los diagramas de persistencia correspondientes con estas dos familias.

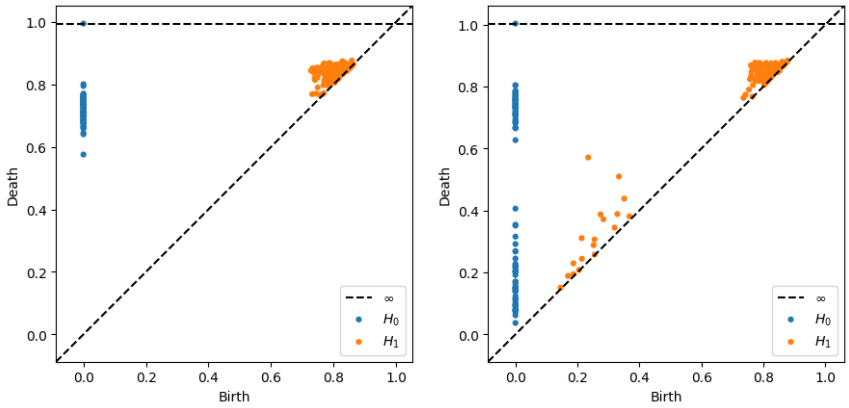

Figura 13: Ejemplos de diagramas de persistencia de la familias simuladas 1 (izquierda) y 2 (derecha). Las dos familias resultan en diagramas de persistencia distintos en ambas dimensiones de homología, e igualmente similares a los diagramas obtenidos con el dataset del IMPAC (Comparar con Sección 7.3).

6.5. Métricas de evaluación de modelos

Dada una variable X de dos clases 0 y 1 (llamadas clase negativa y clase positiva, respectivamente), un predictor, y un conjunto sobre el cual predecir las clases, para cada instancia x tenemos una predicción $\hat{x}$. Las cuatro posibles combinaciones de verdad y predicción se resumen en el siguiente cuadro de doble entrada, con sus respectivos nombres en las celdas:

- Accuracy: Es la proporción de aciertos sobre todas las estimaciones. No distingue entre aciertos de la clase positiva o aciertos de la clase negativa. En términos del cuadro de doble entrada, se calcularía así: (tp + tn) / (tp + tn + fp + fn)
- Precision: La precisión es la fracción de verdaderos positivos sobre todos los positivos estimados, tp / (tp + fp) según el cuadro. Lo que intenta capturar esta métrica es qué tan precisa es la predicción de la clase positiva. En general se codifican los problemas para que la clase positiva sea la clase de interés, y un predictor (más) preciso es un predictor que cumple con

		Clase predicha	
		1	0
Clase verdadera	1	verdaderos positivos (tp)	falsos positivos (fp)
	0	falsos negativos (fn)	verdaderos negativos (tn)

sus predicciones de la clase positiva (por sobre otro predictor). Respondería a la pregunta "¿De todos los que predije como positivos, cuántos realmente lo son?"

- Recall: El Recall también se concentra en las predicciones que resultaron ser verdaderos positivos, pero mirando su relación con la clase positiva real. Esto es: tp / (tp + fn). Responde a la pregunta: "de todos los que efectivamente son de la clase positiva, ¿a cuántos acerté?"

- AUC: El área bajo la curva (*area under the curve*, AUC por sus siglas en inglés) es una medida que es útil cuando el predictor predice *probabilidades de clase* en vez de las clases en sí. Estas probabilidades tienen que ser convertidas de alguna forma en clases predichas, y para eso se necesita un punto de corte. Por ejemplo: a todas las instancias que se les asigne una probabilidad mayor o igual al 0.7 de pertenecer a la clase positiva, se las clasifica como esa clase. Al fijar el 0.7 (llamado "punto de corte"), las clases predichas pueden evaluarse en términos de las medidas anteriores. Para todos los puntos de corte posibles entre el 0 y el 1, entonces, se evalúan dos medidas sobre las clases resultantes: el recall por un lado, y otra medida llamada *false positive rate*[16], por el otro. A cada punto de corte se le asocia un punto en el plano (false positive rate)×(recall), y se toma el área debajo de la curva que los une. Esta medida tiene un mínimo de 0 y un máximo de 1. Un modelo es considerado mejor que otro en términos de esta medida si su área bajo la curva asociada es mayor. Un modelo aleatorio se espera tenga una AUC de 0,5.

[16]Se calcula haciendo fp / (fp + tn). Busca medir la tasa de falsos positivos, sería la respuesta a la pregunta "¿qué fracción de instancias de la verdadera clase negativa clasifiqué como positivas?"

6.6. Modelos con Diagramas

Para poder estimar la utilidad de los diagramas de persistencia en el diagnóstico de los sujetos con TEA, se buscó comparar su performance con la vectorización de las matrices de correlación, que es la práctica estándar en predicción de autismo. Contando con los datos anatómicos de los sujetos del IMPAC además de sus datos funcionales, se incorporaron estos datos también en la predicción.

Para cada una de las parcelaciones disponibles, y para cada método de cálculo de las matrices de similaridad (correlación, correlación parcial, y tangente), se estimó un modelo de Random Forest y un modelo de Gradient Boosting; utilizando en primer lugar los datos anatómicos en conjunto con la vectorización de las matrices de correlación, y en segundo lugar los datos anatómicos en conjunto con la vectorización de los diagramas de persistencia. En total estos modelos descriptos son 84: 7 parcelaciones, 3 métodos, 2 modelos, y 2 procesamientos de datos (matrices de correlación directamente y diagramas de persistencia vectorizados).

El objetivo de estos modelos es comparar las performances de ambas fuentes de datos en condiciones de igualdad. A su vez, se computaron dos modelos más (un Random Forest y un Gradient Boosting, con las mismas especificaciones, descriptas en la sección de Configuración Experimental), pero solamente con los datos anatómicos de los sujetos, a modo de control.

6.7. Modelos Kernel SVM

Los modelos de Support Vector Machine son modelos estándares en la literatura de aprendizaje automático. Son modelos que buscan el hiperplano de mejor separación entre las clases dado un conjunto de datos etiquetados. Generalmente estos modelos utilizan también un *kernel*: una función que computa la cercanía entre dos puntos luego de proyectarlos en un espacio generalmente de dimensión mayor y generalmente no lineal. Este kernel permite a los modelos de SVM atacar problemas de clasificación no lineales, implementando una transformación no lineal que permita una separación lineal en el espacio transformado. Se requiere también para facilitar los cómputos que para dos puntos cualesquiera del conjunto de datos x_i y x_j, la función de kernel cumpla que $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, donde $\cdot$ representa el producto interno estándar en $\mathbb{R}$, para alguna función univariada ϕ.

Gracias a los aportes de Mathieu Carriére creando la librería Scikit-TDA [17], es posible utilizar la implementación de kernels que toman como input a diagramas de persistencia directamente, y computan la separación en el marco de un Support Vector Machine estándar. En este trabajo se utilizaron dos de estos kernels: el que computa la distancia Wasserstein y el kernel persistente de escala espacial (Ver Carrière et al. (2017); Kusano et al. (2017), respectivamente). Se corrieron entonces los modelos de kernel-SVM para investigar la potencialidad de clasificación de los diagramas sin el intermediario de la vectorización, entrenándolos y testeándolos utilizando el mismo pipeline de 8 Fold Cross Validation que el resto de los modelos expuestos.

6.8. Modelos comparados IMPAC

De los modelos de mejor performance en la competencia IMPAC[18] se tomaron por un lado el modelo estándar (referenciado como "starting-kit"), y por el otro el modelo de "MK". Se eligieron dos modelos en vez de todos porque el resto presentó problemas de ejecución, ya sean fallas en el código o un tiempo de cómputo requerido demasiado alto. Se los corrió sobre el conjunto de entrenamiento, nuevamente utilizando 8-fold cross validation y reportando las métricas de accuracy, AUC, recall, y precision. Luego, se modificaron los códigos para utilizar las vectorizaciones de los diagramas de persistencia donde originalmente estaban las matrices de correlación vectorizadas. Las modificaciones se hicieron manteniendo la estructura original de los modelos lo mayor posible, minimizando el imapcto del reemplazo para poder hacer una comparación fidedigna.

El modelo Starting Kit toma a la parcelación MSDL y vectoriza las proyecciones tangenciales de las matrices de covarianza obtenidas de los fMRIs. Utiliza también las variables anatómicas. Para cada uno de estos dos conjuntos de variables ajusta una regresión logística y predice sus probabilidades, con las que ajusta una nueva logística que las toma como variables de entrada y define la predicción final. En su versión modificada se añade una tercera regresión logística al primer grupo que toma como dato a la vectorización de los diagramas de persistencia. Las probabilidades estimadas con ésta se juntan con las otras para conformar la entrada del modelo logístico final. El modelo MK también descompone tangencialmente pero con la parcelación de Craddock Scorr, y luego repite el mismo proceso. La versión modificada añade nuevamente un tercer clasificador logístico que toma como entrada a las vectorizaciones de los diagramas.

[17] https://github.com/MathieuCarriere/sklearn-tda

[18] pueden encontrarse los códigos en `https://github.com/ramp-kits/autism/tree/best_submissions/submissions`

7. Experimentos y resultados

7.1. Configuración experimental

Los modelos de Random Forest utilizaron los parámetros estándares de la librería ScikitLearn 0.24.2, excepto los estimadores, en cuyo caso se utlizaron 300. Los parámetros entonces son:

- n_estimators=300
- criterion="gini"
- max_depth =None
- min_samples_split=2
- min_samples_leaf=1
- min_weight_fraction_leaf=0.0
- max_features="auto"
- max_leaf_nodes=None
- min_impurity_decrease=0.0
- min_impurity_splitf=None
- oob_scorebool=False
- n_jobs=None
- random_state=None
- verbose=0
- warm_start=False
- class_weight=None
- ccp_alphanon-negative=0.0
- max_samples=None

Los modelos de Gradient Boosting utilizaron el paquete XGBoost 1.5.0, con los parámetros estándares, a saber:

- eta =0.3 (learning rate)
- gamma =0 (min split loss)
- max_depth =6
- min_child_weight =1
- max_delta_step =0
- subsample =1
- sampling_method = uniform
- colsample_bytree=1
- colsample_bylevel=1
- colsample_bynode =1
- lambda=1 (reg lambda)
- alpha =0
- tree_method = auto
- sketch_eps =0.03
- scale_pos_weight=1
- updater = grow_colmaker
- refresh_leaf =1
- grow_policy= depthwise
- max_leaves=0
- max_bin=256
- predictor=auto

Los modelos fueron entrenados utilizando 8-fold cross validation, y se reporta de cada modelo la accuracy, precision, AUC, y recall promedios, junto a sus respectivas varianzas. Todos los modelos fueron corridos en Python 3.6.9, en Google Colab.

7.2. Proyecciones de diagramas vectorizados

Pueden apreciarse en la figura 14 las dos primeras componentes principales de la vectorización de los diagramas de persistencia en el conjunto de entrenamiento con los varios caminos de preprocesamiento, combinando distancias dinámicas y coseno con las matrices de correlación, correlación parcial, y tangentes. En rojo se ve la clase negativa y en azul la clase positiva. Los porcentajes explicados por estas dos primeras componentes principales se resumen en el Cuadro 2.

No se observa ninguna separación apartente entre clases, y esto es independiente del tipo de matriz y distancia que se utilice. También se observa que la distancia utilizada no afecta demasiado la proyección, y es interesante notar el gran porcentaje de variabilidad explicada en las dos primeras componentes de ambos casos con correlación y correlación parcial.

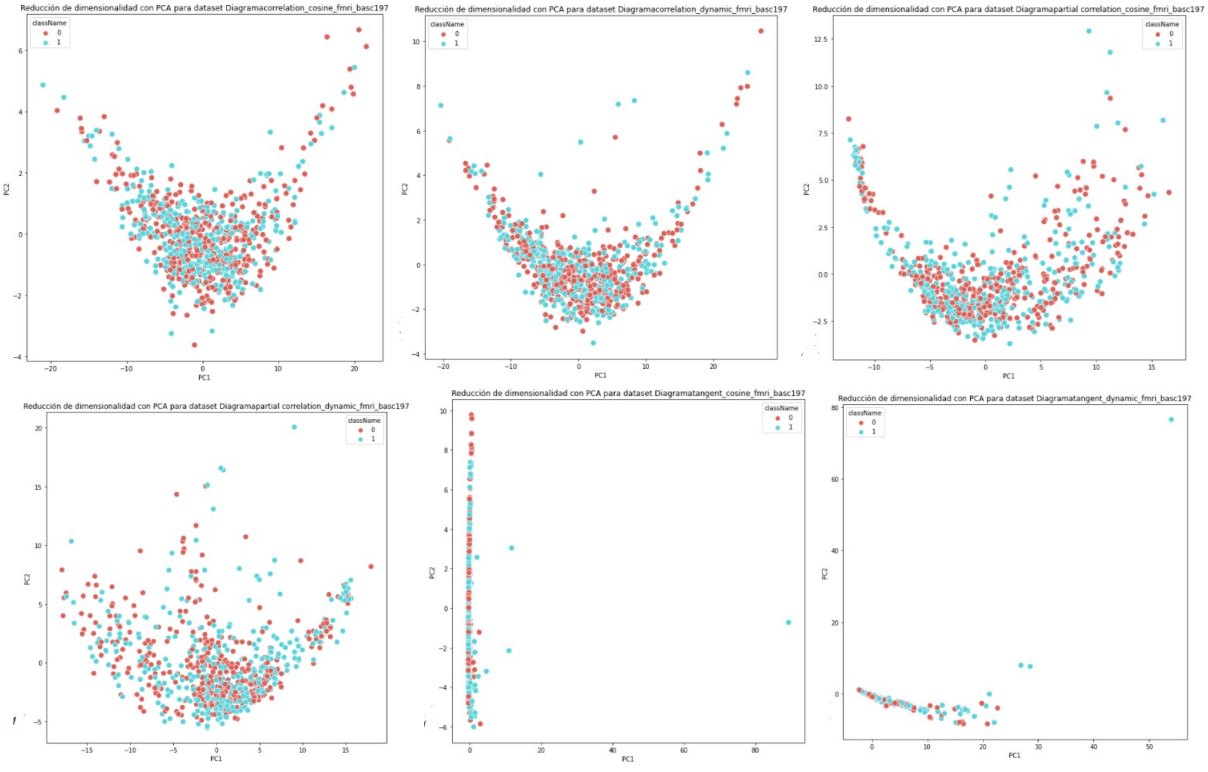

Figura 14: Proyecciones de diagramas vectorizados con las matrices de distancia computadas con distancia coseno en columna izquierda, y en la derecha distancia dinámica; y computadas con correlación, correlación parcial, y proyección tangencial arriba, al centro, y abajo, respectivamente. No se observa una clara separación de clases en ninguna instancia.

		Cómputo		
		Correlación	**Corr Parcial**	**Tangente**
Distancia	**Coseno**	95 %	90 %	69 %
	Dinámica	94 %	87 %	66 %

Cuadro 2: Varianza explicada de los dos primeros componentes principales según la matriz en la vectorización de diagramas. Vemos que la mayoría de la variabilidad está concentrada en estas dos.

7.3. Exploración de proyecciones

Para poder explorar la falta de separación de clases que se observó en las proyecciones de los diagramas, se vectorizaron los diagramas del conjunto de entrenamiento de la parcelación BASC-197, y se seleccionaron puntos de cada clase que estuvieran claramente separados (a simple vista). Pueden verse los puntos y la subselección en la Figura 15. Estos puntos se utilizaron para verificar que el proceso de vectorización utilizado no era el causal de esta falta de separación. Se pudieron ver también los diagramas asociados a estos puntos y corroborar que efectivamente son distintos. Esto puede apreciarse en las figuras 16 y 17.

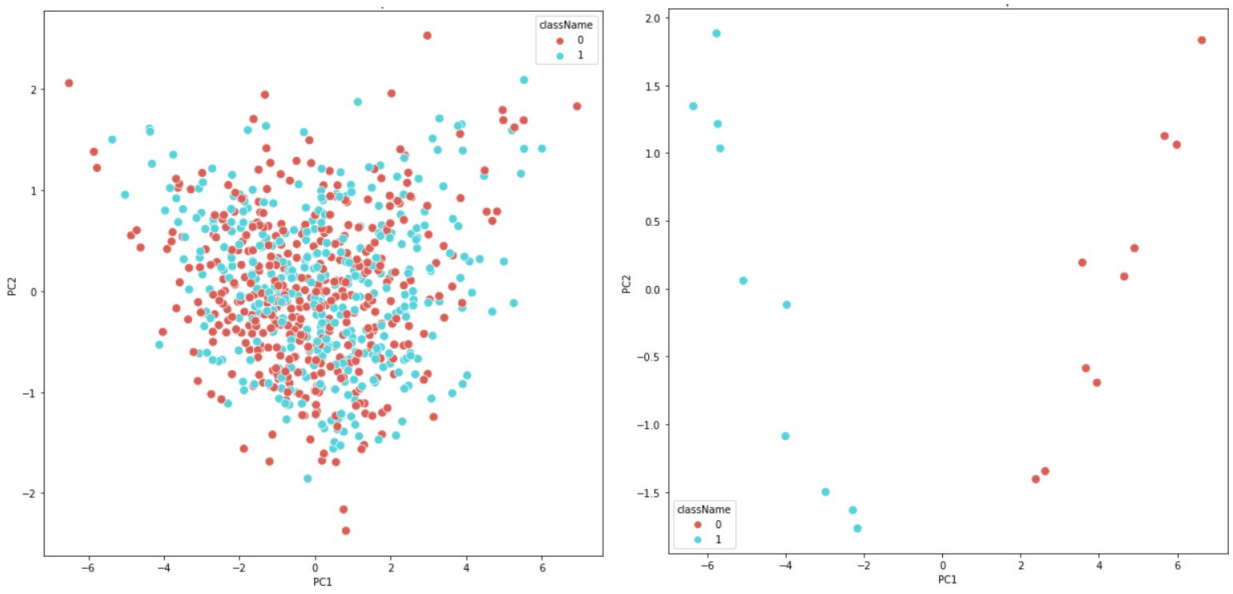

Figura 15: Izquierda: Proyección sobre las dos primeras componentes principales de la vectorización de diagramas de persistencia, coloreados por clase. Derecha: puntos elegidos como subconjunto separable

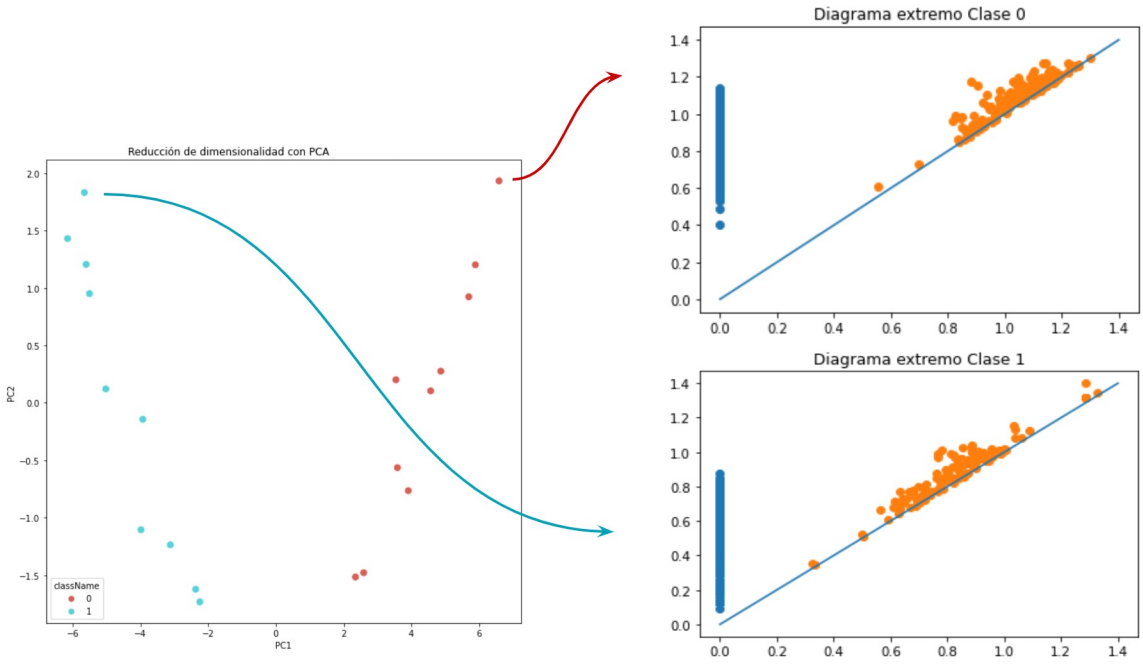

Figura 16: Diagramas de persistencia asociados a los puntos señalados en la vectorización. Vemos dos diagramas distintos que se corresponden con puntos alejados en la proyección. Las clases difieren en la posición de los puntos de homología 0 (azules) y en la homología 1 (naranjas).

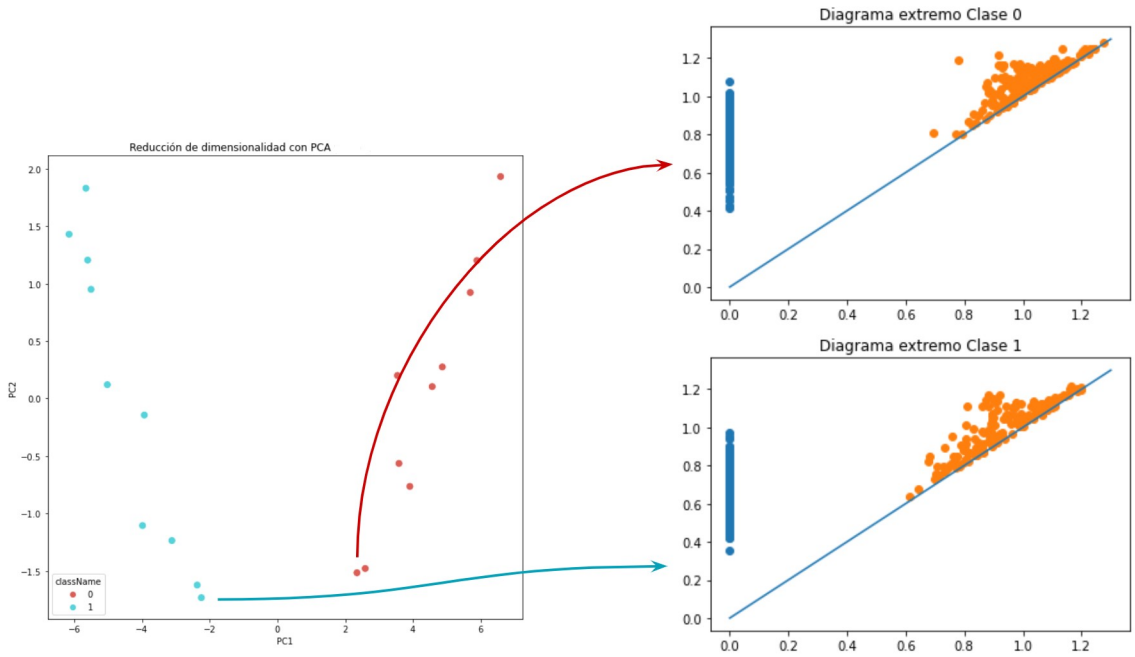

Figura 17: Más diagramas de persistencia asociados a los puntos señalados en la vectorización. Vemos en este caso que estos dos diagramas se corresponden con puntos más cercanos en la vectorización, y son más similares entre sí que los de la Figura anterior.

Una prueba más que se hizo con estos puntos aislados es entrenar un modelo de SVM directamente con los diagramas asociados. Se utilizaron estas instancias como input a modelos de SVM con los siguientes kernels exclusivos para diagramas de persistencia:

- Sliced Wasserstein Kernel
- Persistence Weighted Gaussian Kernel
- Persistence ScaleSpace Kernel
- Persistence Fisher Kernel

Los tres primeros tuvieron una Accuracy de 1 sobre este subconjunto de diagramas, mientras que el último solo alcanzó una accuracy del 50 %. Esto se considera evidencia de que los diagramas

cuyas vectorizaciones están claramente separadas también constituyen en sí mismos un espacio separable, y de que el problema hallado en la separación de las proyecciones de los diagramas vectorizados es más profundo que simplemente un error de cómputo en la vectorización.

Por supuesto, la falta de separación a simple vista evaluando las primeras dos componentes principales no es evidencia suficiente para afirmar que un espacio no es separable. De todas formas, en la figura 15 (izq.), las primeras dos componentes principales capturan el 84 % de la variabilidad de los datos. En general, se observa que todas las proyecciones de las vectorizaciones concentran su variabilidad en las primeras componentes principales, en contraste por ejemplo con las proyecciones de las matrices de correlación (vectorizando su triángulo superior).

7.4. Proyecciones de matrices vectorizados

En el mismo espíritu que la sección anterior se exploraron también las proyecciones de las matrices de conectividad vectorizadas. La separación visual de los casos por clase sigue sin ser aparente, pero disminuye considerablemente el porcentaje de variabilidad contenido en las dos primeras componentes principales. En el caso de las matrices de correlación y correlación parcial, tenemos comprendido solamente un 0,08 % y 0,02 %, respectivamente; mientras que en la descomposición tangencial se captura el 45 % en las primeras dos (Ver Figura 18).

7.5. Prueba de Concepto

En las señales simuladas con las dos familias de series de tiempo se puede visualizar una clara separación de los casos en las primeras dos componentes principales. Esta separación se muestra independiente del preprocesamiento de la matriz de correlaciones y de su conversión a distancia. Pueden verse tres ejemplos de estas proyecciones en la Figura 19.

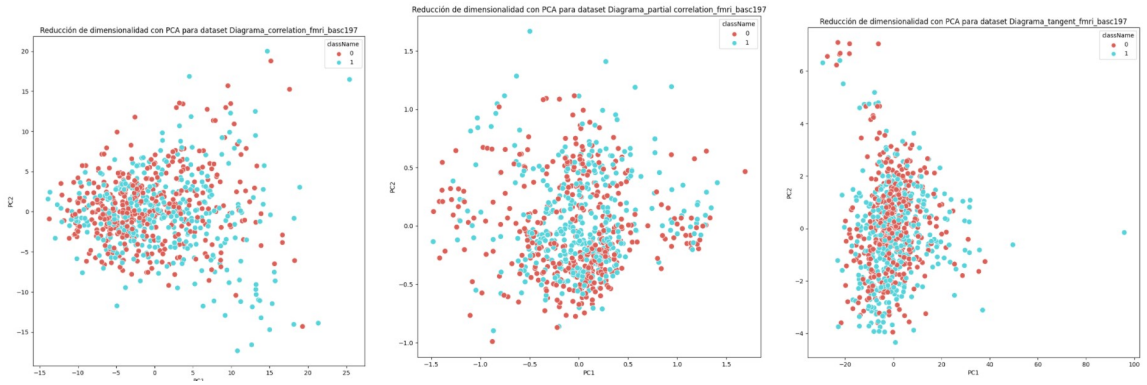

Figura 18: Proyecciones de las primeras dos componentes principales de las matrices de conectividad sobre el conjunto de entrenamiento. Las técnicas de preprocesamiento son: con matriz de correlaciones y distancia coseno, con correlaciones parciales y distancia coseno, y con descomposición tangencial y distancia coseno (izquierda, centro, y derecha, resp.). En rojo se ve la clase negativa, y en azul la positiva (presencia de TEA). En estos casos así como en las proyecciones de los diagramas no se observa una clara separación de clases, pero en estas dos primeras componentes está capturada considerablemente menos variabilidad.

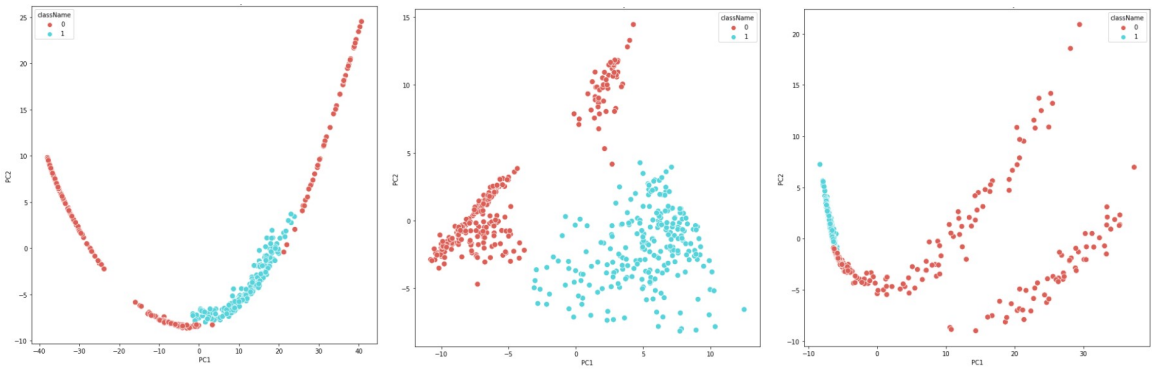

Figura 19: Proyecciones de las primeras dos componentes principales de los diagramas de persistencia correspondientes a las series generadas. Las técnicas de preprocesamiento son: con matriz de correlaciones y distancia coseno, con correlaciones parciales y distancia coseno, y con descomposición tangencial y distancia dinámica (izquierda, centro, y derecha, resp.). En rojo se ve la clase negativa, y en azul la positiva (presencia de TEA). En los tres casos la separación de clases es visualmente evidente, validando el proceso de vectorización de los diagramas y descartándolo como causal de la falta de separación evidente en los datos del IMPAC.

7.6. Entropías

La ventaja de calcular las entropías para los niveles de homología elegidos es que al solo tener dos dimensiones de homología, a cada diagrama le corresponde solamene un par de números,

que pueden ser graficados en un plano directamente, prescindiendo de técnicas como el PCA. No se observa una separación en términos de la entropía de los diagramas tampoco. En ninguna de las dimensiones individualmente, ni en su interacción en el plano. Puede observarse en la figura 20 dos ejemplos, ambos para la parcelación BASC-197, donde la falta de separación es clara.

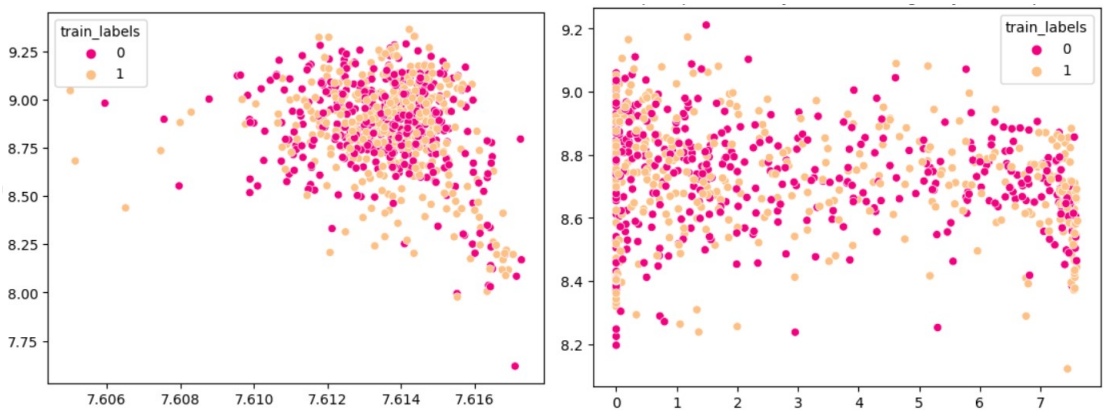

Figura 20: Scatterplot de entropías de las homologías de dimensión 0 en el eje x, y la entropía de las homologías de dimensión 1 en el eje y. Izquierda: entropía de diagramas calculados con correlación parcial y distancia dinámica. Derecha: entropías de diagramas calculados con descomposición tangencial y distancia coseno. La entropía no resulta ser una métrica útil para separar las clases tampoco.

7.6.1. Entropía como filtro

Se siguió el trabajo de Atienza et al. (2016) para implementar un algoritmo de filtrado de puntos relevantes de un diagrama de persistencia según la información que aportaran al diagrama. El filtrado de los diagramas de persistencia no resultó en una diferencia perceptible en las proyecciones visualizadas. En términos del diagrama en sí, resultó equivalente a remover todos los puntos a una distancia fija de la diagonal. Se obvió entonces el uso de este filtro, considerando también que el preprocesamiento de los diagramas de persistencia utilizado tenía la opción de remover puntos a una distancia fija de la diagonal, y/o quedarse solamente con los primeros k puntos más lejanos a la diagonal.

7.7. Curvas de Betti

En las curvas de Betti no se observa una separación por clase. Se encuentra también una amplia variación en las curvas de distintos diagramas. Como puede verse en la Figura 21, las clases están casi completamente solapadas. Se muestran solamente dos ejemplos de BASC-197, pero el patrón se repite para el resto de las técnicas de preprocesamiento. Las bandas que pueden verse rodeando a la línea media capturan al primer desvío estándar de los datos. Entonces no solamente las curvas de Betti no muestran una separación entre clases, sino que sus varianzas también ocupan grandes rangos, haciendo que la curva de Betti media por clase sea un estimador poco preciso de los valores reales según el valor del parámetro. Si las clases estuvieran más separadas tendría sentido realizar algún tipo de análisis estadístico paramétrico con el objetivo de estimar la probabilidad por clase de un diagrama basándose en su relación con las curvas medias por clase, pero se descartó esta posibilidad en vista de los gráficos expuestos.

Se calcularon también las curvas de Betti sobre el subconjunto de casos seleccionados a mano del dataset IMPAC expuestos en las figuras 16 y 17. Éstas curvas pueden verse en la figura 22. Así como el conjunto en su totalidad, este subconjunto de datos tampoco presenta una separación aparente entre las curvas de Betti por clase para ninguno de sus dos niveles de homología. Se exponen a modo comparativo las curvas de Betti de la figura 23, correspondientes a los casos simulados en la prueba de concepto, que sí muestran un mayor grado de separación.

7.8. Modelos IMPAC

En la siguiente tabla 4 pueden verse los resultados de los modelos IMPAC comparados con sus contrapartes que incorporan la vectorización de los diagramas de persistencia en el pipeline de predicción. Se puede apreciar que la incorporación de las features topológicas no empeora las métricas, pero tampoco las mejora. Los resultados son prácticamente iguales indicando que la incorporación de estas variables extra no aportó información útil para los modelos.

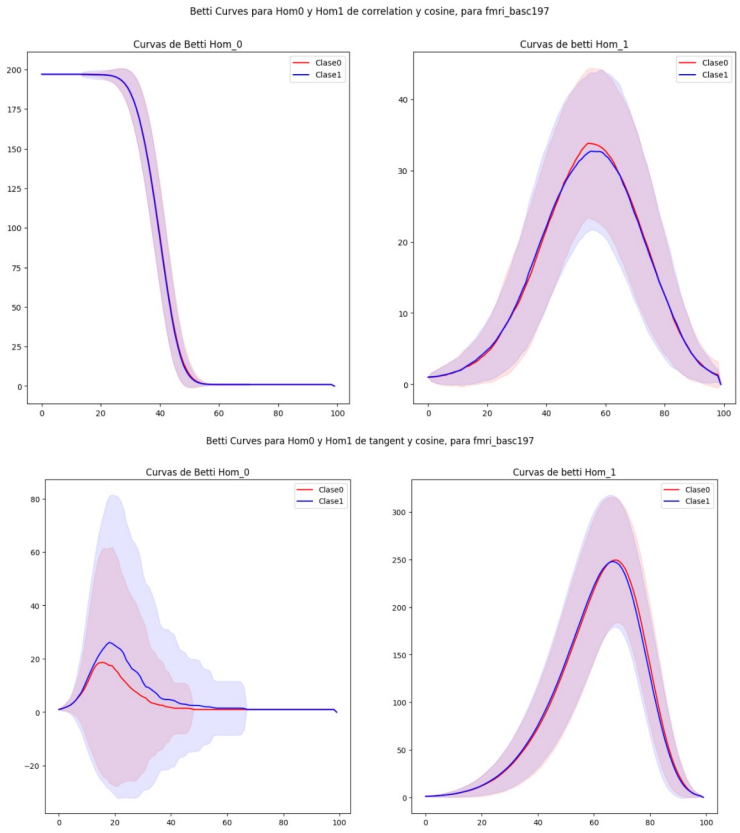

Figura 21: Curvas de Betti para BASC197 con correlación y distancia coseno (arriba), y descomposición tangencial y distancia coseno (abajo). En cada caso, se ven las curvas de la homología de dimensión 0 a la izquierda, y 1 a la derecha. La clase negativa está graficada en rojo, y la positiva en azul. El color violeta se debe a la superposición de las clases. Estas curvas no muestran ninguna separación de clases.

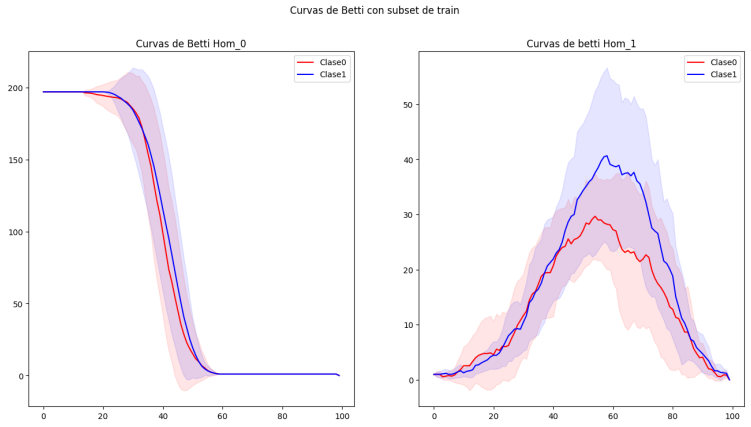

Figura 22: Se pueden ver las curvas de Betti de homologías 0 y 1 (izquierda y derecha, resp.) de el subconjunto de puntos elegidos del dataset de entrenamiento. Son similares a las del conjunto total y tampoco muestran una separación evidente de clases; con la diferencia que al ser curvas promedio de menos casos, los valores son más irregulares.

Modelo	**Versión**	**AUC**	**Accuracy**	**Precision**	**Recall**
Starting Kit	**Original**	0.649 (0.023)	0.606 (0.023)	0.597 (0.028)	0.546 (0.050)
	Modificado	0.650 (0.023)	0.602 (0.022)	0.591 (0.026)	0.556 (0.073)
MK	**Original**	0.732 (0.021)	0.664 (0.018)	0.659 (0.021)	0.616 (0.035)
	Modificado	0.732 (0.021)	0.663 (0.023)	0.654 (0.018)	0.627 (0.059)

Cuadro 4: Resultados de los modelos originales de la competencia IMPAC y sus versiones modificadas. Entre paréntesis se ven los desvíos estándares del 8FCV. Las versiones originales y modificadas tienen una performance prácticamente idénticas.

7.9. Modelos propios

Se buscó correr modelos que pudieran comparar la performance predictiva de los diagramas de persistencia con respecto a las matrices de correlación. Se observa que los clasificadores que utilizan las matrices directamente oscilan en performance con una varianza mayor a aquellos que utilizan los diagramas de persistencia. Si bien en algunos casos las matrices tienen una performance

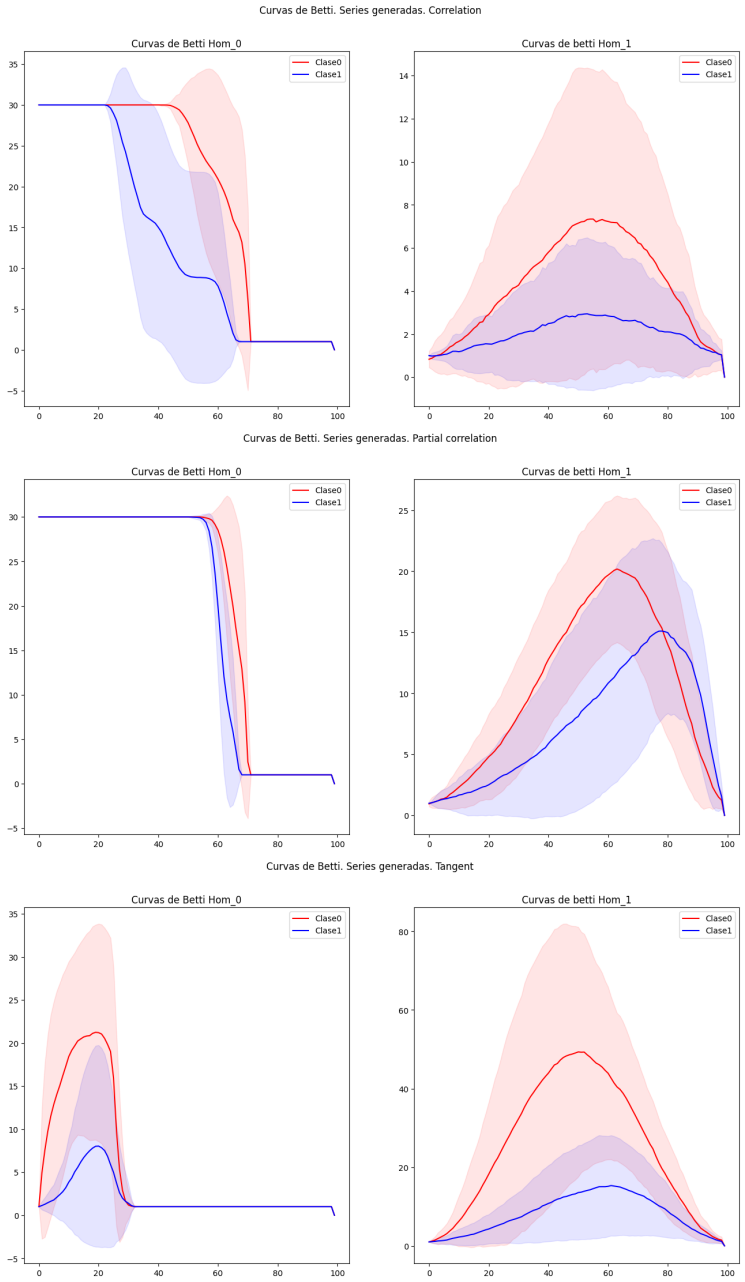

Figura 23: Curvas de Betti de homologias 0 y 1 (izquierda y derecha, resp.) de las secuencias generadas como prueba de concepto. Se ven, en orden, las generadas a partir de las matrices de correlación, correlación parcial, y proyección tangencial. En estos casos hay una separación visualmente evidente de las clases fabricadas, a diferencia de las curvas de Betti hechas con casos reales del IMPAC.

superior, no son la mayoría, y la variabilidad que tienen los resultados hace atractivo al uso de diagramas de persistencia en la predicción.

Independientemente de la medida de performance, los diagramas de persistencia obtuvieron métricas con menor varianza, y con performance media igual o muy similar a las matrices. Recordemos que hay dos fuentes de variabilidad en este caso. Por un lado, tenemos la varianza de las ocho estimaciones del 8-fold cross validation. Por el otro, está la varianza entre todas las estimaciones promedio de cada estimador. En el primer gráfico de la Figura 24, vemos cómo se distribuyen las métricas de las estimaciones puntuales. Allí se ve que en AUC, Accuracy, y Precision, las estimaciones promedio de cada 8-FCV varían menos que las estimaciones promedio de los modelos que utilizan las matrices de correlación. En el caso del recall vemos que las estimaciones promedio tienen una alta varianza en ambos casos, con una pequeña mejora en el recall en caso de los modelos "propios". Sin embargo, incluso para el caso del Recall, puede verse en la Figura 28, que la mayoría de los modelos tiene un Recall cercano al 0,5 para ambas familias de modelos.

Tenemos entonces dos vistas diferentes: En un menor nivel de abstracción tenemos a las Figuras 25, 26, 27, y 28. Aquí, se ven los 42 modelos para cada familia: 42 modelos que utilizan información anatómica y directamente la matriz (ya sea correlación, correlación parcial, o tangente), y 42 modelos que utilizan la información anatómica junto con los diagramas de persistencia vectorizados. Los modelos se juntaron lado a lado para facilitar la comparación: se ven uno al lado del otro los modelos de matrices y de diagramas, que utilizan la misma parcelación, misma forma de obtener la matriz, y mismo modelo (se utilizó la distancia coseno para convertir a la matriz de similaridad en una matriz de distancia en todos los casos). Para cada combinación entonces de modelo, matriz, parcelación, y tipo de dato, se grafica el promedio de las métricas junto con líneas superiores e inferiores cuyo largo es un desvío estándar. En estos gráficos también pueden apreciarse unos pequeños puntos grises: estos aparecen en aquellos casos en los que los intervalos de $+/-$ un desvío estándar de las métricas no se solapan. Por encima de estos puntos grises, los puntos amarillos indican cuando esta diferencia es a favor de los modelos que utilizan diagramas de persistencia.

El modelo que que utiliza la parcelación BASC197 con correlación parcial y estimador XGBoost de diagramas de persistencia es superior en todas las métricas a su análogo con la matriz. En las cuatro métricas los intervalos de un desvío estándar no se solapan y la diferencia es a favor de los diagramas de persistencia. En los mismos gráficos, los dos puntos de más a la derecha en amarillo corresponden a dos modelos que utilizan **solamente las características anatómicas del sujeto**, estimando con un Random Forest y un XGBoost respectivamente. Podemos ver que

los modelos de diagramas de persistencia no aportan realmente a la clasificación, al menos no de una forma significativa, porque su performance es siempre igual a la de los modelos puramente anatómicos. Comparando ahora con los modelos que utilizan las matrices, vemos que la mejoría en la clasificación es pequeña, y se contrarresta con una mayor variabilidad que los que utilizan solamente caraterísticas anatómicas.

7.10. Modelos sin anatomía

7.10.1. SVM sobre diagramas

Se puede apreciar en la tabla 6 que los modelos que utilizaron directamente un kernel sobre los diagramas de persistencia y sin aporte de datos anatómicos clasifican como clasificaría en esperanza una asignación aleatoria.

Modelo	**AUC**	**Accuracy**	**Precision**	**Recall**
Wasserstein	0.491	0.493	0.474	0.503
Kernel	(0.052)	(0.036)	(0.042)	(0.333)
Persistence Scale	0.504	0.491	0.411	0.515
Space Kernel	(0.065)	(0.018)	(0.080)	(0.483)

Cuadro 6: Métricas sobre los modelos Kernel SVM. Los modelos tienen una performance aleatoria en la tarea de clasificación en toas las métricas y para ambos Kernels.

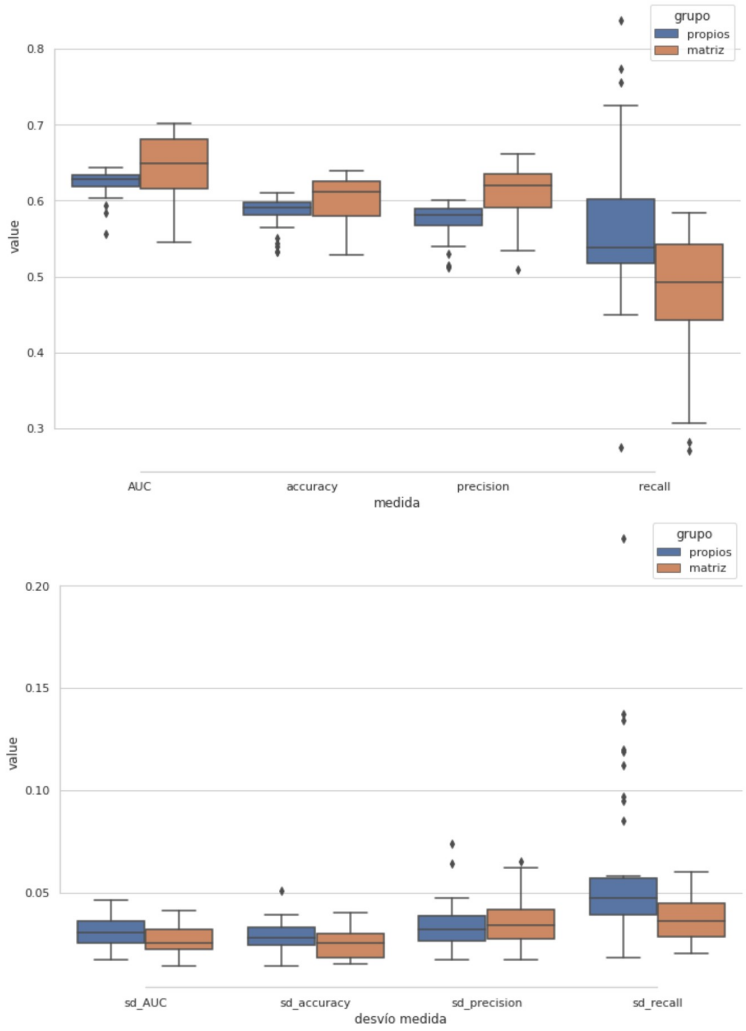

Figura 24: Resumen de medidas de performance (arriba) y sus desvíos estándares del 8FCV (abajo), comparando los modelos que utilizan matrices con los modelos que utilizan diagramas de persistencia. Vemos que los resultados puntuales de los modelos de matrices de correlación son mejores en promedio que los de features topológicas, pero a costa de una mayor variabilidad a través de especificaciones de modelos.

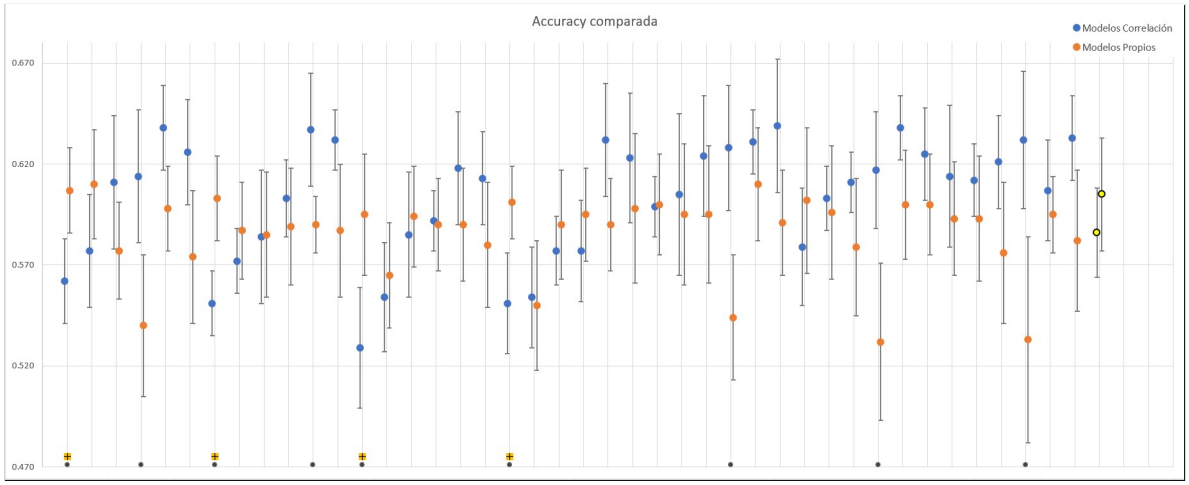

Figura 25: Accuracy comparada de todos los modelos corridos. En azul los modelos con matrices de conectividad, en naranja los modelos de diagramas de persistencia. Los puntos grises demarcan los modelos donde hay diferencia entre los intervalos de desvíos estándares, y los puntos amarillos por encima aquellos casos donde la diferencia es a favor de los modelos topológicos. Los modelos de las matrices de correlación oscilan más que sus contrapartes topológicas, y se ve una varianza similar. De los nueve intervalos que no se solapan, en cuatro son mejores los modelos topológicos.

7.10.2. Random Forest sobre diagramas y sobre matrices de conectividad

Se corrieron también modelos sin incluir las características anatómicas de los sujetos. El objetivo es poder separar el aporte de cada input distinto y entender su efecto en la predicción total. En particular, poder comparar lado a lado el aporte de los diagramas de persistencia por sobre las matrices de conectividad. Por supuesto, esta separación no es perfecta: nada quita que al juntar los datos anatómicos con alguna de estas dos fuentes mencionadas la superficie a discriminar se facilite o dificulte, ya que en general las performances de los modelos en su totalidad no son aditivas con respecto a las de sus componentes por separado. De todas formas, resulta interesante ver cómo se comportan estos modelos en relación a aquellos que utilizan también anatomía.

En vista de que en los modelos de la sección “Modelos propios” no presentaron una variabilidad discernible según el tipo de modelo que se utilizó, se eligió correr solamente modelos Random Forest con las mismas especificaciones ya mencionadas, con el objetivo de simplificar el análisis de los resultados. En otro afán simplificador, se utilizaron solamente la correlación parcial y la descomposición tangencial (omitiendo la correlación) para el cálculo de las matrices de conectividad.

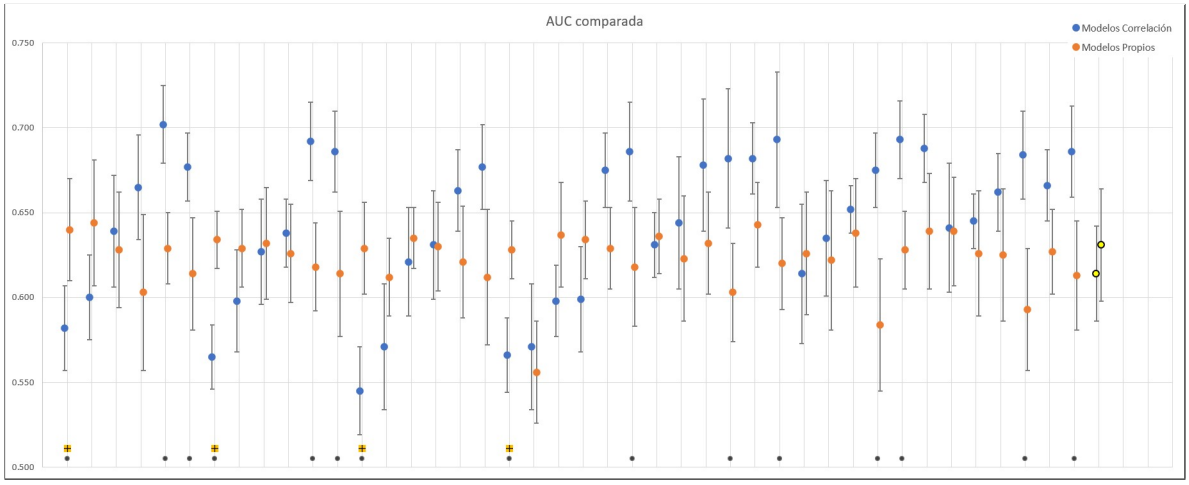

Figura 26: AUC comparada de todos los modelos corridos. En azul los modelos con matrices de conectividad, en naranja los modelos de diagramas de persistencia. Los puntos grises demarcan los modelos donde hay diferencia entre los intervalos de desvíos estándares, y los puntos amarillos por encima aquellos casos donde la diferencia es a favor de los modelos topológicos. Se ve claramente una performance estable de los modelos topológicos independientemente de la especificación, pero la mayoría de las diferencias considerables entre los modelos es a favor del uso de matrices de correlación.

Tenemos entonces 28 modelos en total: usando matrices de conectividad directamente o sus diagramas generados; calculando la conectividad con correlaciones parciales o descomposiciones tangenciales; y utilizando las siete parcelaciones a disposición. Se pueden apreciar estos modelos comparados lado a lado en las Figuras 30, 29, 31, y 32.

Estos resultados dilucidan los observados en los modelos que sí utilizan anatomía. En ese caso, se observa que los modelos que incorporan los diagramas de persistencia tienen un comportamiento estable en sus métricas que es muy similar al comportamiento de los modelos que utilizan solo la anatomía. Asimismo, los modelos de correlación evidencian una mayor variabilidad y una performance superadora en algunos casos. En los modelos que utilizan solamente los diagramas de persistencia las métricas son prácticamente equivalentes a aquellas que se obtendrían de una elección aleatoria entre dos clases (50 %), mientras que las matrices de correlación por sí solas alcanzan métricas un poco mejores, como un 60 % de accuracy. Se destaca de todas formas el modelo propio de XGB con la parcelación BASC197 y correlación parcial para calcular la matriz de distancias,

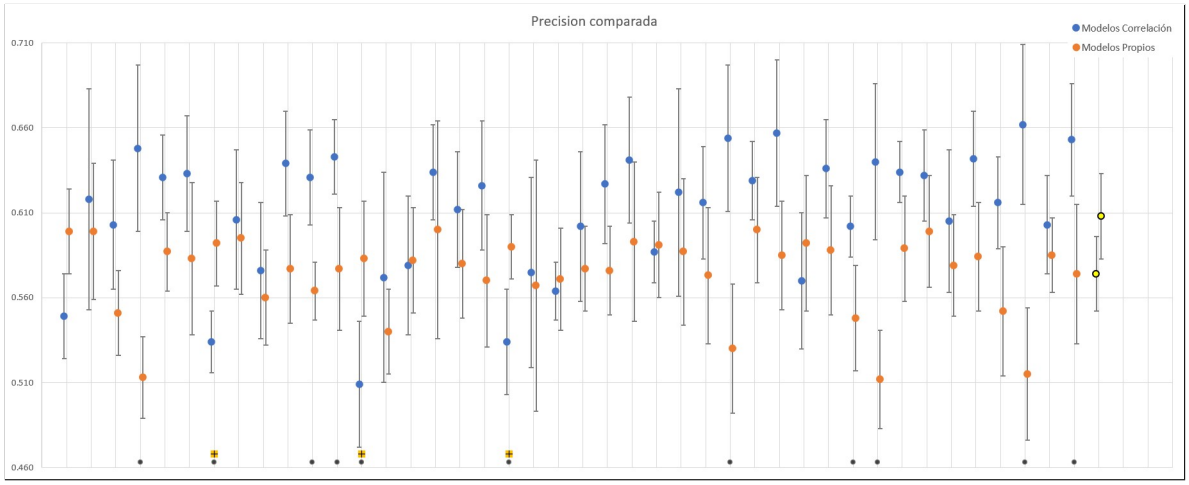

Figura 27: Precision comparada de todos los modelos corridos. En azul los modelos con matrices de conectividad, en naranja los modelos de diagramas de persistencia. Los puntos grises demarcan los modelos donde hay diferencia entre los intervalos de desvíos estándares, y los puntos amarillos por encima aquellos casos donde la diferencia es a favor de los modelos topológicos. Se observan performances estables de los modelos de topología, pero ninguno mejor que los modelos que utilizan solo anatomía (puntos extremos derechos en amarillo).

que en todos los casos superó a su contraparte de vectorización de la matriz de correlación. Este resultado sin embargo no parecería ser estadísticamente significativo, y esta tendencia se revierte en los modelos sin anatomia.

Recordemos también que las proyecciones de los diagramas de persistencia concentran la mayoría de su variabilidad en las primeras componentes principales. Tenemos entonces un conjunto de features que varían principalmente en dos o tres dimensiones, y que tienen una performance aleatoria con respecto a la etiqueta de clase. Esto explicaría que los modelos que incorporan estas featureas a la anatomía no mejoran la clasificación, porque agregarle los diagramas de persistencia vectorizados sería equivalente a agregarle ruido aleatorio en unas pocas dimensiones. En cambio, el uso directo de las matrices de conectividad vectorizadas sí puede alcanzar mayores métricas cuando se las junta con las features anatómicas ya que, aunque poca, aportan información nueva a la clasificación (por más que también aumenten la variabilidad de todos los estimadores). Los resultados de estos modelos y su aparente independencia de la parcelación utilizada confirman las advertencias ejercidas por Rathore et al. (2019), y queda claro que un 70 % de accuracy es el máximo resultado que puede obtenerse del dataset IMPAC.

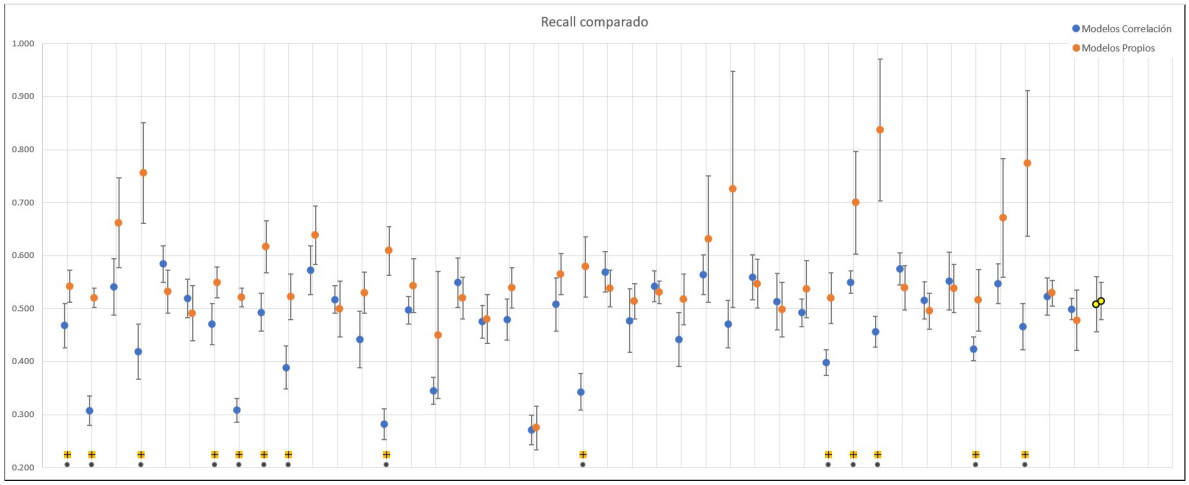

Figura 28: Recall comparada de todos los modelos corridos. En azul los modelos con matrices de conectividad, en naranja los modelos de diagramas de persistencia. Los puntos grises demarcan los modelos donde hay diferencia entre los intervalos de desvíos estándares, y los puntos amarillos por encima aquellos casos donde la diferencia es a favor de los modelos topológicos. La mayoría de los modelos oscila poco, con algunas excepciones positivas en los modelos de topología. Donde los intervalos de varianza no se solapan la diferencia es siempre a favor de los modelos con topología.

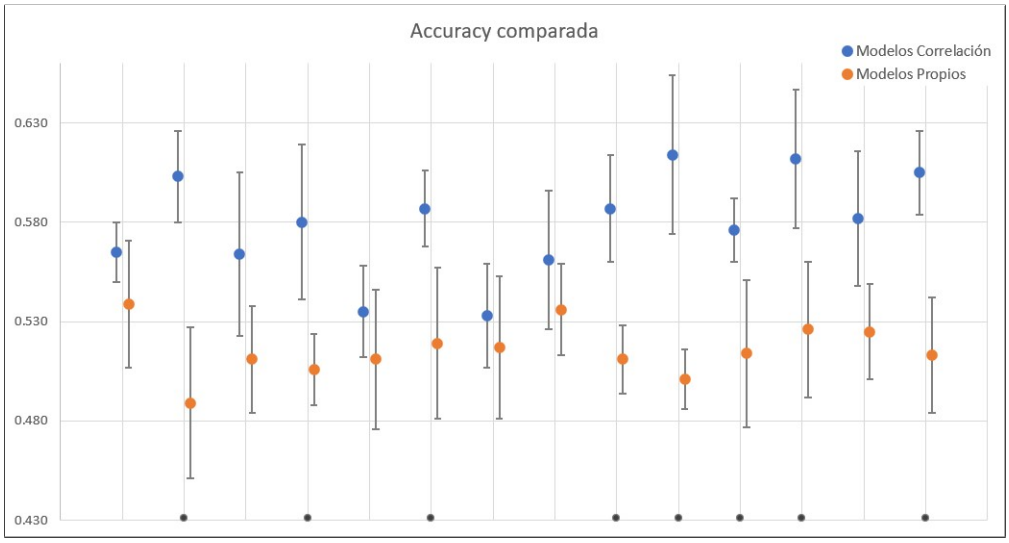

Figura 29: Accuracy comparada del subconjunto de modelos corridos nuevamente pero sin anatomía. Se observa que la performance de los modelos propios es prácticamente aleatoria, mientras que los modelos de correlación tienen una accuracy al rededor del 60 %

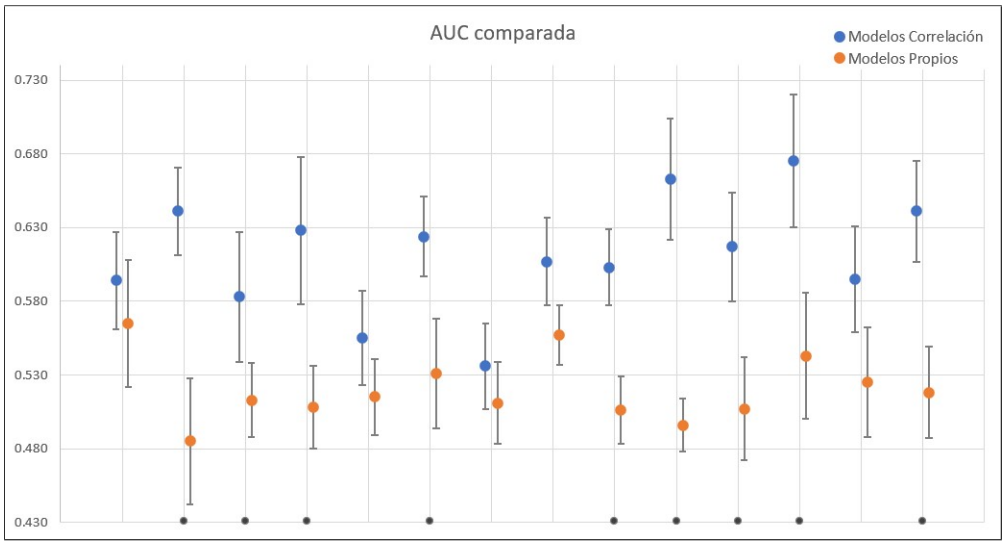

Figura 30: AUC comparada del subconjunto de modelos corridos nuevamente pero sin anatomía. La AUC es considerablemente mejor en los modelos de correlación comparado con los modelos propios, que oscilan en el 50 % como una asignación aleatoria.

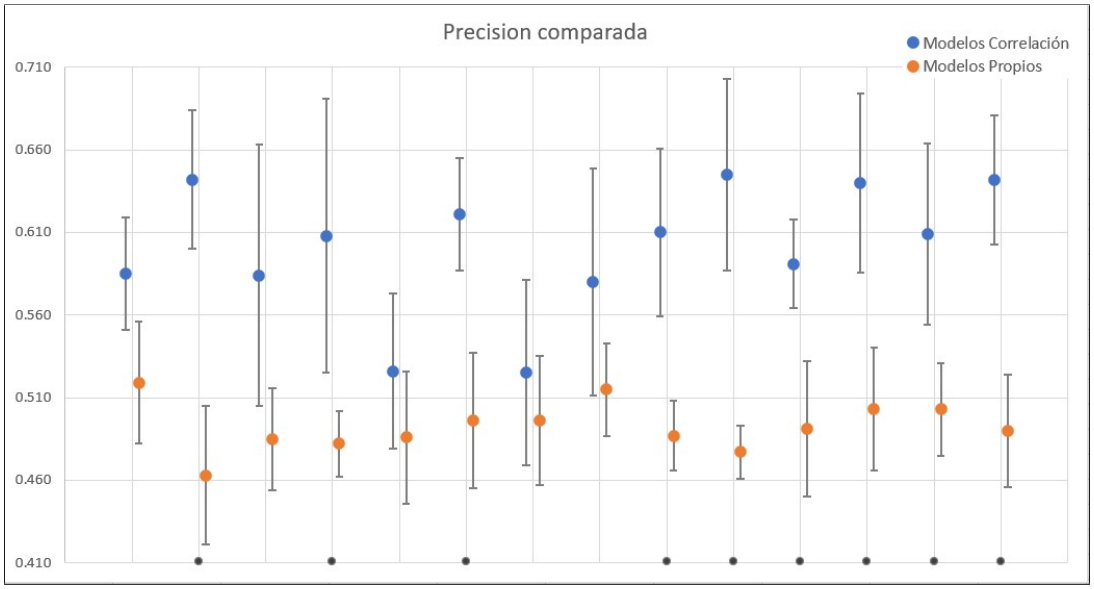

Figura 31: Precision comparada del subconjunto de modelos corridos nuevamente pero sin anatomía. Se observa nuevamente una performance casi aleatoria de los modelos propios comparado con los modelos de correlación que alcanzan en algunos casos 70 % de Precision.

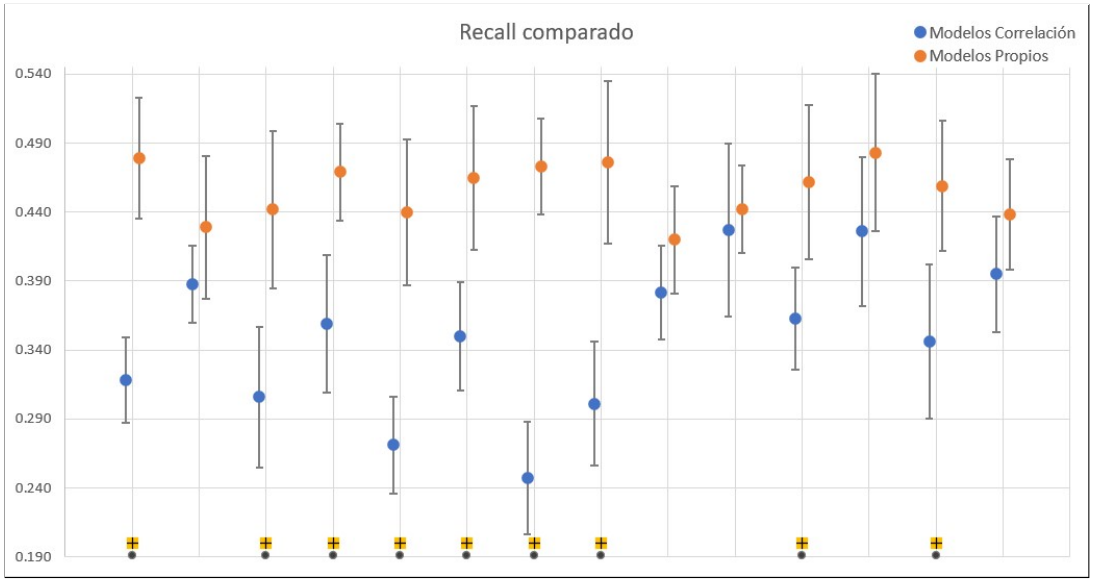

Figura 32: Recall comparada del subconjunto de modelos corridos nuevamente pero sin anatomía. A diferencia de las métricas anteriores, el recall es más alto en los modelos propios que en la correlción, pero siguen teniendo una performance prácticamente aleatoria.

8. Discusión

En este trabajo se exploraron las posibilidades de diagnóstico de autismo utilizando resúmenes de las características topológicas de los conectomas de sujetos. Esto se hizo en comparación con el estado del arte y también buscando complementarlo. Se investigaron en general las técnicas de homología persistente para esta tarea evaluando su efectividad en predicción y separación de clases para el dateset del IMPAC, para una muestra del dataset ABIDE, y en casos generados como prueba de concepto.

Se entiende que el objetivo ideal sería diagnosticar con certeza la presencia de trastornos del espectro autista, pero la falta de un resultado positivo en la incorporación de las características topológicas a los clasificadores no resultara desesperanzador. Hay mucho que aprender de estos resultados. Como mencionan en Caputi et al. (2021), en el contexto de la esquizofrenia parecería que la diferencia entre las clases que puede obtenerse desde los fMRIs proviene de características topográficas en vez de topológicas. Esto es: los niveles de conectividad cambian de forma minúscula pero perceptible entre las clases pero este cambio no es suficiente como para generar un efecto medible en la conformación topológica de los conectomas de los sujetos. Los diagramas de persistencia son robustos ante el ruido: los resultados de estabilidad de los diagramas aseguran que la diferencia

entre dos diagramas distintos está acotada superiormente por la diferencia entre las funciones que los generaron. En este caso, es esa misma robustez que al parecer borra las diferencias entre clases. La verdadera *feature*, se ve, se encuentra en aquellos recovecos de los grafos de conectividad que la homología pasa como ruido.

Habiendo atacado el problema desde todos los ángulos y exhaustivamente probado combinaciones de modelos predictivos, podemos ver que efectivamente no hay una diferencia en la conformación topológica de los conectomas según su diagnóstico de autismo. Simplemente la información no se encuentra allí. En una tarea tan compleja como es la detección del autismo, poder comprender qué caminos *no* seguir es un aporte también valioso para la búsqueda. Por encima de esto, incluso en los casos en donde se consigue un resultado positivo (usando directamente las matrices de conectividades y features anatómicas), solamente se llega a un 70 % de predicciones acertadas. Esto confirma algo que no es nuevo: los trastornos del espectro autista son difíciles de diagnosticar de forma certera, así como son difíciles de comprender.

Hay problemas que pueden también adjudicársele a los datos. Como mencionan en Rathore et al. (2019), las ventanas de obtención de las resonancias magnéticas son cortas (menores a 10 minutos), y son resonancias en estado de reposo. Esto deja abiertas las dudas sobre qué pasaría si se contara con mediciones de mayor longitud o si se trabajaran con resonancias con estímulo. No resultaría extremo pensar que las personas con TEA sí muestren diferencias en la conformación topológica de sus conectomas frente un estímulo con respecto a un grupo control. Asimismo, la etiqueta binaria de presencia o ausencia de autismo deja mucho que desear: como lo indica el nombre (TEA), el autismo existe en un espectro, y verlo de forma binaria oscurece el análisis y potencialmente borra la capacidad discriminadora de las features. Sim embargo, el problema de esta crítica a la etiqueta es más grande que el dataset IMPAC en sí, porque no es una falla en la obtención de datos sino que, al momento, no hay una escala que pueda capturar los distintos niveles de autismo: no tiene sentido hablar de ser "autista al 0.8". A su vez, el mismo intento de diagnóstico de autismo por medio de resonancias magnéticas funcionales acarrea el supuesto de que allí se encuentra el autismo, y no queda claro todavía si la causa de los trastornos autistas es neurológica. En todo caso, estos resultados apuntarían a una causalidad conjunta, donde parte del autismo afecta o proviene de diferencias en la conectividad cerebral, y el resto de otra fuente (posiblemente social).

Mientras el análisis topológico de datos sigue encontrando su nicho en los diagnósticos y análisis neurológicos, la predicción del autismo sigue siendo un bastión por conquistar. Queda pendiente explorar la homología dirigida, y resonancias funcionales que no sean de reposo. Seguramente

los avances venideros en recolección de datos y los desarrollos constantes en los modelos predictivos puedan desembocar en un futuro en un diagnóstico útil de los Trastornos del espectro Autista, contribuyendo a la detección y accionar terapeutico temprano y a mejorar la vida de quienes sean parte del Espectro.

9. Anexo

9.1. Resultados Modelos con Diagramas de Persistencia

A continuación se presentan los resultados de los modelos que utilizan features anatómicas y las vectorizaciones de los diagramas de persistencia. Las abreviaturas HOCP y CSM refieren a las parcelaciones de Harvard Oxford y Craddock Scorr Mean, respectivamente, mientras que B064, B122, y B197 refieren a las parcelaciones BASC-64, BASC-122, y BASC-197.

ATLAS	**TIPO**	**MODELO**	**ROC-AUC**	**accuracy**	**precision**	**recall**
B122	PCOR	XGB	0.640 (.030)	0.607 (0.021)	0.599 (0.025)	0.542 (0.030)
B123	PCOR	RF	0.644 (0.037)	0.610 (0.027)	0.599 (0.040)	0.520 (0.018)
B124	TAN	XGB	0.628 (0.034)	0.577 (0.024)	0.551 (0.025)	0.662 (0.085)
B125	TAN	RF	0.603 (0.046)	0.540 (0.032)	0.513 (0.024)	0.756 (0.095)
B126	CORR	XGB	0.629 (.021)	0.598 (0.021)	0.587 (0.023)	0.532 (0.041)
B127	CORR	RF	0.614 (0.033)	0.574 (0.033)	0.583 (0.045)	0.491 (0.052)
B197	PCOR	XGB	0.634 (.017)	0.603 (0.021)	0.592 (0.025)	0.549 (0.029)
B198	PCOR	RF	0.629 (0.023)	0.600 (0.023)	0.595 (0.033)	0.521 (0.018)
B199	TAN	XGB	0.632 (.033)	0.585 (0.031)	0.560 (0.028)	0.617 (0.049)
B200	TAN	RF	0.626 (0.029)	0.589 (0.029)	0.577 (0.032)	0.522 (0.043)
B201	CORR	XGB	0.618 (0.026)	0.590 (0.014)	0.564 (0.017)	0.638 (0.055)
B202	CORR	RF	0.614 (0.037)	0.587 (0.033)	0.577 (0.036)	0.499 (0.053)

Table 7 continued from previous page

ATLAS	TIPO	MODELO	ROC-AUC	accuracy	precision	recall
CSM	PCOR	XGB	0.629 (0.027)	0.595 (0.030)	0.583 (0.034)	0.53 (0.039)
CSM	PCOR	RF	0.612 (0.023)	0.565 (0.026)	0.540 (0.025)	0.609 (0.046)
CSM	TAN	XGB	0.635 (0.018)	0.594 (0.025)	0.582 (0.031)	0.543 (0.051)
CSM	TAN	RF	0.630 (0.026)	0.59 (0.023)	0.60 (0.064)	0.45 (0.120)
CSM	CORR	XGB	0.621 (0.033)	0.59 (0.028)	0.58 (0.032)	0.52 (0.039)
CSM	CORR	RF	0.612 (0.040)	0.58 (0.031)	0.57 (0.039)	0.48 (0.046)
P2011	PCOR	XGB	0.628 (0.017)	0.601 (0.018)	0.590 (0.019)	0.539 (0.038)
P2012	PCOR	RF	0.556 (0.030)	0.550 (0.032)	0.567 (0.074)	0.275 (0.041)
P2013	TAN	XGB	0.637 (0.031)	0.590 (0.027)	0.571 (0.030)	0.565 (0.039)
P2014	TAN	RF	0.634 (0.023)	0.595 (0.023)	0.577 (0.025)	0.579 (0.057)
P2015	CORR	XGB	0.629 (0.024)	0.590 (0.023)	0.576 (0.026)	0.538 (0.034)
P2016	CORR	RF	0.618 (0.035)	0.598 (0.037)	0.593 (0.047)	0.514 (0.033)
MSDL	PCOR	XGB	0.636 (0.022)	0.600 (0.025)	0.591 (0.031)	0.531 (0.021)
MSDL	PCOR	RF	0.623 (0.037)	0.595 (0.035)	0.587 (0.043)	0.517 (0.048)
MSDL	TAN	XGB	0.632 (0.030)	0.595 (0.034)	0.573 (0.040)	0.631 (0.119)
MSDL	TAN	RF	0.603 (0.029)	0.544 (0.031)	0.530 (0.038)	0.725 (0.223)

Table 7 continued from previous page

ATLAS	TIPO	MODELO	ROC-AUC	accuracy	precision	recall
MSDL	CORR	XGB	0.643 (0.025)	0.610 (0.028)	0.600 (0.031)	0.547 (0.046)
MSDL	CORR	RF	0.620 (0.027)	0.591 (0.026)	0.585 (0.032)	0.498 (0.051)
B064	PCOR	XGB	0.626 (0.036)	0.602 (0.036)	0.592 (0.040)	0.537 (0.054)
B065	PCOR	RF	0.622 (0.041)	0.596 (0.033)	0.588 (0.038)	0.520 (0.048)
B066	TAN	XGB	0.638 (0.032)	0.579 (0.034)	0.548 (0.031)	0.700 (0.097)
B067	TAN	RF	0.584 (0.039)	0.532 (0.039)	0.512 (0.029)	0.837 (0.134)
B068	CORR	XGB	0.628 (0.023)	0.600 (0.027)	0.589 (0.031)	0.539 (0.042)
B069	CORR	RF	0.639 (0.034)	0.600 (0.025)	0.599 (0.033)	0.495 (0.034)
HOCP	PCOR	XGB	0.639 (0.032)	0.593 (0.028)	0.579 (0.030)	0.538 (0.045)
HOCP	PCOR	RF	0.626 (0.037)	0.593 (0.031)	0.584 (0.032)	0.516 (0.058)
HOCP	TAN	XGB	0.625 (0.039)	0.576 (0.035)	0.552 (0.038)	0.671 (0.112)
HOCP	TAN	RF	0.593 (0.036)	0.533 (0.051)	0.515 (0.039)	0.774 (0.137)
HOCP	CORR	XGB	0.627 (0.025)	0.595 (0.019)	0.585 (0.022)	0.529 (0.024)
HOCP	CORR	RF	0.613 (0.032)	0.582 (0.035)	0.574 (0.041)	0.478 (0.057)

9.2. Resultados Modelos Matrices solo vectorizadas

A continuación se presentan los resultados de los modelos que utilizan features anatómicas y las vectorizaciones de las matrices de correlación. Las abreviaturas HOCP y CSM refieren a las parcelaciones de Harvard Oxford y Craddock Scorr Mean, respectivamente, mientras que B064, B122, y B197 refieren a las parcelaciones BASC-64, BASC-122, y BASC-197.

ATLAS	**TIPO**	**MODELO**	**ROC-AUC**	**accuracy**	**precision**	**recall**
B122	PCOR	XGB	0.582 (0.025)	0.562 (0.021)	0.549 (0.025)	0.468 (0.042)
B123	PCOR	RF	0.600 (0.025)	0.577 (0.028)	0.618 (0.065)	0.307 (0.028)
B124	TAN	XGB	0.639 (0.033)	0.611 (0.033)	0.603 (0.038)	0.541 (0.053)
B125	TAN	RF	0.665 (0.031)	0.614 (0.033)	0.648 (0.049)	0.419 (0.052)
B126	CORR	XGB	0.702 (0.023)	0.638 (0.021)	0.631 (0.025)	0.584 (0.035)
B127	CORR	RF	0.677 (0.020)	0.626 (0.026)	0.633 (0.034)	0.519 (0.036)
B197	PCOR	XGB	0.565 (0.019)	0.551 (0.016)	0.534 (0.018)	0.471 (0.039)
B198	PCOR	RF	0.598 (0.030)	0.572 (0.016)	0.606 (0.041)	0.308 (0.022)
B199	TAN	XGB	0.627 (0.031)	0.584 (0.033)	0.576 (0.040)	0.493 (0.036)
B200	TAN	RF	0.638 (0.020)	0.603 (0.019)	0.639 (0.031)	0.389 (0.041)
B201	CORR	XGB	0.692 (0.023)	0.637 (0.028)	0.631 (0.028)	0.573 (0.046)
B202	CORR	RF	0.686 (0.024)	0.632 (0.015)	0.643 (0.022)	0.517 (0.026)
CSM	PCOR	XGB	0.545 (0.026)	0.529 (0.030)	0.509 (0.037)	0.442 (0.053)

Table 8 continued from previous page

ATLAS	TIPO	MODELO	ROC-AUC	accuracy	precision	recall
CSM	PCOR	RF	0.571 (0.037)	0.554 (0.027)	0.572 (0.062)	0.282 (0.029)
CSM	TAN	XGB	0.621 (0.032)	0.585 (0.031)	0.579 (0.041)	0.497 (0.026)
CSM	TAN	RF	0.631 (0.032)	0.592 (0.015)	0.634 (0.028)	0.345 (0.025)
CSM	CORR	XGB	0.663 (0.024)	0.618 (0.028)	0.612 (0.034)	0.549 (0.047)
CSM	CORR	RF	0.677 (0.025)	0.613 (0.023)	0.626 (0.038)	0.475 (0.031)
P2011	PCOR	XGB	0.566 (0.022)	0.551 (0.025)	0.534 (0.031)	0.479 (0.039)
P2012	PCOR	RF	0.571 (0.037)	0.554 (0.025)	0.575 (0.056)	0.271 (0.028)
P2013	TAN	XGB	0.598 (0.021)	0.577 (0.017)	0.564 (0.017)	0.508 (0.050)
P2014	TAN	RF	0.599 (0.031)	0.577 (0.025)	0.602 (0.044)	0.343 (0.035)
P2015	CORR	XGB	0.675 (0.022)	0.632 (0.028)	0.627 (0.035)	0.569 (0.038)
P2016	CORR	RF	0.686 (0.029)	0.623 (0.032)	0.641 (0.037)	0.477 (0.060)
MSDL	PCOR	XGB	0.631 (0.019)	0.599 (0.015)	0.587 (0.018)	0.542 (0.029)
MSDL	PCOR	RF	0.644 (0.039)	0.605 (0.040)	0.622 (0.061)	0.442 (0.051)
MSDL	TAN	XGB	0.678 (0.039)	0.624 (0.030)	0.616 (0.033)	0.564 (0.037)
MSDL	TAN	RF	0.682 (0.041)	0.628 (0.031)	0.654 (0.043)	0.471 (0.045)
MSDL	CORR	XGB	0.682 (0.021)	0.631 (0.016)	0.629 (0.023)	0.559 (0.042)

Table 8 continued from previous page

ATLAS	TIPO	MODELO	ROC-AUC	accuracy	precision	recall
MSDL	CORR	RF	0.693 (0.040)	0.639 (0.033)	0.657 (0.043)	0.513 (0.053)
B064	PCOR	XGB	0.614 (0.041)	0.579 (0.029)	0.570 (0.040)	0.492 (0.026)
B065	PCOR	RF	0.635 (0.034)	0.603 (0.016)	0.636 (0.029)	0.398 (0.024)
B066	TAN	XGB	0.652 (0.014)	0.611 (0.015)	0.602 (0.018)	0.550 (0.021)
B067	TAN	RF	0.675 (0.022)	0.617 (0.029)	0.640 (0.046)	0.456 (0.029)
B068	CORR	XGB	0.693 (0.023)	0.638 (0.016)	0.634 (0.018)	0.575 (0.030)
B069	CORR	RF	0.688 (0.020)	0.625 (0.023)	0.632 (0.027)	0.516 (0.035)
HOCP	PCOR	XGB	0.641 (0.038)	0.614 (0.035)	0.605 (0.042)	0.552 (0.054)
HOCP	PCOR	RF	0.645 (0.016)	0.612 (0.018)	0.642 (0.028)	0.424 (0.022)
HOCP	TAN	XGB	0.662 (0.023)	0.621 (0.023)	0.616 (0.027)	0.547 (0.038)
HOCP	TAN	RF	0.684 (0.026)	0.632 (0.034)	0.662 (0.047)	0.466 (0.044)
HOCP	CORR	XGB	0.666 (0.021)	0.607 (0.025)	0.603 (0.029)	0.523 (0.035)
HOCP	CORR	RF	0.686 (0.027)	0.633 (0.021)	0.653 (0.033)	0.499 (0.020)

9.3. Resultados Modelos sin anatomía

A continuación se presentan los resultados de los modelos de Random Forest que no utilizan datos anatómicos de los sujetos. La columna "fuente" es "DIAG" si se utilizaron los diagramas de persistencia vectorizados y "MAT" si se utilizaron las matrices de conectividad aplanadas. Las

abreviaturas HOCP y CSM refieren a las parcelaciones de Harvard Oxford y Craddock Scorr Mean, respectivamente, mientras que B064, B122, y B197 refieren a las parcelaciones BASC-64, BASC-122, y BASC-197.

fuente	**parcelacion**	**matriz**	**modelo**	**ROC-AUC**	**accuracy**	**precision**	**recall**
DIAG	B122	PCOR	RF	0.565 (0.043)	0.539 (0.032)	0.519 (0.037)	0.479 (0.044)
DIAG	B122	TAN	RF	0.485 (0.043)	0.489 (0.038)	0.463 (0.042)	0.429 (0.052)
DIAG	B197	PRCOR	RF	0.513 (0.025)	0.511 (0.027)	0.485 (0.031)	0.442 (0.057)
DIAG	B197	TAN	RF	0.508 (0.028)	0.506 (0.018)	0.482 (0.020)	0.469 (0.035)
DIAG	CSM	PCOR	RF	0.515 (0.026)	0.511 (0.035)	0.486 (0.040)	0.440 (0.053)
DIAG	CSM	TAN	RF	0.531 (0.037)	0.519 (0.038)	0.496 (0.041)	0.465 (0.052)
DIAG	P2011	PRCOR	RF	0.511 (0.028)	0.517 (0.036)	0.496 (0.039)	0.473 (0.035)
DIAG	P2011	TAN	RF	0.557 (0.020)	0.536 (0.023)	0.515 (0.028)	0.476 (0.059)
DIAG	MSDL	PCOR	RF	0.506 (0.023)	0.511 (0.017)	0.487 (0.021)	0.420 (0.039)
DIAG	MSDL	TAN	RF	0.496 (0.018)	0.501 (0.015)	0.477 (0.016)	0.442 (0.032)
DIAG	B064	PRCOR	RF	0.507 (0.035)	0.514 (0.037)	0.491 (0.041)	0.462 (0.056)
DIAG	B064	TAN	RF	0.543 (0.043)	0.526 (0.034)	0.503 (0.037)	0.483 (0.057)
DIAG	HOCP	PCOR	RF	0.525 (0.037)	0.525 (0.024)	0.503 (0.028)	0.459 (0.047)
DIAG	HOCP	TAN	RF	0.518 (0.031)	0.513 (0.029)	0.490 (0.034)	0.438 (0.040)
MAT	B122	PRCOR	RF	0.594 (0.033)	0.565 (0.015)	0.585 (0.034)	0.318 (0.031)

MAT	B122	TAN	RF	0.641 (0.030)	0.603 (0.023)	0.642 (0.042)	0.388 (0.028)
MAT	B197	PCOR	RF	0.583 (0.044)	0.564 (0.041)	0.584 (0.079)	0.306 (0.051)
MAT	B197	TAN	RF	0.628 (0.050)	0.580 (0.039)	0.608 (0.083)	0.359 (0.050)
MAT	CSM	PRCOR	RF	0.555 (0.032)	0.535 (0.023)	0.526 (0.047)	0.271 (0.035)
MAT	CSM	TAN	RF	0.624 (0.027)	0.587 (0.019)	0.621 (0.034)	0.350 (0.039)
MAT	P2011	PCOR	RF	0.536 (0.029)	0.533 (0.026)	0.525 (0.056)	0.247 (0.041)
MAT	P2011	TAN	RF	0.607 (0.030)	0.561 (0.035)	0.580 (0.069)	0.301 (0.045)
MAT	MSDL	PRCOR	RF	0.603 (0.026)	0.587 (0.027)	0.610 (0.051)	0.382 (0.034)
MAT	MSDL	TAN	RF	0.663 (0.041)	0.614 (0.040)	0.645 (0.058)	0.427 (0.063)
MAT	B064	PCOR	RF	0.617 (0.037)	0.576 (0.016)	0.591 (0.027)	0.363 (0.037)
MAT	B064	TAN	RF	0.675 (0.045)	0.612 (0.035)	0.640 (0.054)	0.426 (0.054)
MAT	HOCP	PRCOR	RF	0.595 (0.036)	0.582 (0.034)	0.609 (0.055)	0.346 (0.056)
MAT	HOCP	TAN	RF	0.641 (0.034)	0.605 (0.021)	0.642 (0.039)	0.395 (0.042)

9.4. Resultados Modelos solo Anatomía

Se presentan los resultados de los modelos que solamente utilizaron features anatómicas.

Modelo	**AUC**	**ACC**	**PREC**	**REC**
anatomy RF	0.631 (0.033)	0.605 (0.028)	0.608 (0.025)	0.514 (0.035)
anatomy XGB	0.614 (0.028)	0.586 (0.022)	0.574 (0.022)	0.508 (0.052)

9.5. Prueba de $\partial^2 = 0$

Proposición 1: La composición de mapas de borde resulta en el mapa nulo.

Demostración:

Se definen mapas $\partial_n : C_n(X) \to C_{n-1}(X)$ por su acción sobre los generadores σ_α, con $\partial_n(\sigma_\alpha) = \sum_{i=0}^{n} \sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]}$ donde $\sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]}$ denota el $(n-1)$-simplejo formado por las caras de σ_α excluyendo el i-ésimo vértice.

Luego $\partial_{n-1} \circ \partial_n(\sigma_\alpha) = \sum_{i=0}^{n} \partial_{n-1}(\sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]})$. Aplicando el segundo operador obtenemos:

$$\partial_{n-1} \circ \partial_n(\sigma_\alpha) = \sum_{i=0}^{n} \left(\sum_{0 \le j < i} \sigma_\alpha|_{[v_0,\ldots,\hat{v}_j,\ldots,\hat{v}_i,\ldots,v_n]} + \sum_{i < j \le n} \sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,\hat{v}_j,\ldots,v_n]} \right)$$

Para cada para (i,j) distintos, $\sigma_\alpha|_{[v_0,\ldots,\hat{v}_j,\ldots,\hat{v}_i,\ldots,v_n]} = \sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,\hat{v}_j,\ldots,v_n]}$. Cada término entonces se encuentra duplicado y, estando en $\mathbb{F}_2$, esto resulta en que todos los terminos tengan el coeficiente cero, entonces $\partial_{n-1} \circ \partial_n(\sigma_\alpha) = 0$

Como cada elemento de $C_n(X)$ es una combinación lineal de generadores y los mapas involucrados son lineales, habiendo establecido la proposición sobre los generadores se sigue que cualquier combinación lineal de éstos también resulta en el mapa nulo bajo el mapa $\partial_{n-1} \circ \partial_n$.

En el caso que no se trabaje sobre $\mathbb{F}_2$, la prueba es similar pero un poco más involucrada en el álgebra. En su generalidad, $\partial_n(\sigma_\alpha)$ se define como $\sum_{i=0}^{n}(-1)^n \sigma_\alpha|_{[v_0,\ldots,\hat{v}_i,\ldots,v_n]}$, pero como en $\mathbb{F}_2$, $(-1)^n \equiv 1$ la definición se simplifica.

Bibliografía

Aktas, M.E., A. E. and Fatmaoui, A. (2019). Persistence homology of networks: methods and applications. *Applied Network Science*, 4.

Atienza, N., Gonzalez-Diaz, R., and Rucco, M. (2016). Separating topological noise from features using persistent entropy.

Bassett, D. and Sporns, O. (2017). A detailed characterization of complex networks using information theory. *Nature Neuroscience*, 20(4502):353—-364.

Bellec, P., Rosa, P., Lyttelton, O., Benali, H., and Evans, A. (2010). Multi-level bootstrap analysis of stable clusters (basc) in resting-state fmri. *NeuroImage*, 51:1126–39.

Berry, E., C. Y. C.-K. J. e. a. (2020). Functional summaries of persistence diagrams. *Appl. and Comput. Topology.*, (4):211—-262.

Byrge, L. and Kennedy, D. P. (2020). Accurate prediction of individual subject identity and task, but not autism diagnosis, from functional connectomes. *Human Brain Mapping*, 41(9):2249–2262.

Caputi, L., Pidnebesna, A., and Hlinka, J. (2021). Promises and pitfalls of topological data analysis for brain connectivity analysis.

Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., and Umeda, Y. (2020). Perslay: A neural network layer for persistence diagrams and new graph topological signatures.

Carrière, M., Cuturi, M., and Oudot, S. (2017). Sliced wasserstein kernel for persistence diagrams.

Christensen, D. L., Braun, K. V. N., Baio, J., Bilder, D., Charles, J., Constantino, J. N., Daniels, J., Durkin, M. S., Fitzgerald, R. T., Kurzius-Spencer, M., Lee, L.-C., Pettygrove, S., Robinson, C., Schulz, E., Wells, C., Wingate, M. S., Zahorodny, W., and Yeargin-Allsopp, M. (2018). Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2012. *MMWR. Surveillance Summaries*, 65(13):1–23.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2006). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120.

Craddock, R., James, G., Holtzheimer, P., Hu, X., and Mayberg, H. (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., and Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fmri. *NeuroImage*, 192:115 – 134.

Edelsbrunner, H. and Harer, J. (2008). Persistent homology—a survey.

Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*.

Ellis, C. T., Lesnick, M., Henselman-Petrusek, G., Keller, B., and Cohen, J. D. (2019). Feasibility of topological data analysis for event-related fmri. *Network neuroscience*, 3:695—-706.

Hong, S.-J., de Wael, R. V., Bethlehem, R. A. I., Lariviere, S., Paquola, C., Valk, S. L., Milham, M. P., Martino, A. D., Margulies, D. S., Smallwood, J., and Bernhardt, B. C. (2019). Atypical functional connectome hierarchy in autism. *Nature Communications*, 10(1).

Kusano, G., Fukumizu, K., and Hiraoka, Y. (2017). Kernel method for persistence diagrams via kernel embedding and weight factor.

Medina, P. S. and Doerge, R. W. (2016). Statistical methods in topological data analysis for complex, high-dimensional data.

Michel, B. (2015). *A Statistical Approach to Topological Data Analysis*. PhD thesis.

Pennec, X., Fillard, P., and Ayache, N. (2006). A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66.

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., , and Petersen, S. E. (2011). Functional network organization of the human brain. *Neuron*, 72:665–678.

Rathore, A., Palande, S., Anderson, J. S., Zielinski, B. A., Fletcher, P. T., and Wang, B. (2019). Autism classification using topological features and deep learning: A cautionary tale. In *Lecture Notes in Computer Science*, pages 736–744. Springer International Publishing.

Stolz, B. J., Emerson, T., Nahkuri, S., Porter, M. A., and Harrington, H. A. (2020). Topological data analysis of task-based fmri data from experiments on schizophrenia.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information Processing in Medical Imaging*, volume 6801 of *Lecture Notes in Computer Science*, pages 562–573, Kaufbeuren, Germany. Gábor Székely, Horst Hahn, Springer.

Venkatesh, M., Jaja, J., and pessoa, L. (2019). Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. *bioRxiv*.