Trabajo Fin de Máster

Cambio en el N-glucoma sérico en pacientes con cáncer

Sergio Antón Fuente

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Cambio en el N-glucoma sérico en pacientes con cáncer
Título en español: Cambio en el N-glucoma sérico en pacientes con cáncer
English title: Changes in serum N-glycome in cancer patients
Modalidad: Modalidad B
Autor/a: Sergio Antón Fuente, Universidad de A Coruña
Director/a: Francisco Gude Sampedro, Complexo Hospitalario Universitario de Santiago de Compostela y Barbanza; Manuel Febrero Bande, Universidad de Santiago
Tutor/a: Jose Luis Otero Ferrer, Biostatech – Advice, Training & Innovation in Biostatistics
Breve resumen del trabajo: El cáncer es una de las enfermedades con mayor impacto a nivel global, siendo vital su diagnóstico precoz. En este trabajo se pretende hacer uso de las técnicas del análisis composicional para dar con clasificadores adecuados a partir de datos del N—glucoma. Adicionalmente se comparan con métodos clásicos que no hacen uso de este enfoque
Recomendaciones:
Otras observaciones:

Don/doña Francisco Gude Sampedro, Adjunto Unidad de Epidemiología e Investigación Clínica de la Complexo Hospitalario Universitario de Santiago de Compostela y Barbanza, don/doña Manuel Febrero Bande, Catedrático de la Universidad de Santiagoy don/doña Jose Luis Otero Ferrer, Asesor senior del Departamento de asesoramiento, formación e I+D+i de Biostatech – Advice, Training & Innovation in Biostatistics, informan que el Trabajo Fin de Máster titulado

Cambio en el N-glucoma sérico en pacientes con cáncer

fue realizado bajo su dirección por don/doña Sergio Antón Fuente para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 14 de julio de 2023.

El/la director/a:
Don/doña Francisco Gude Sampedro

El/la director/a:
Don/doña Manuel Febrero Bande

El/la tutor/a:
Don/doña Jose Luis Otero Ferrer

El/la autor/a:
Don/doña Sergio Antón Fuente

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Agradecer a Francisco Gude y a Óscar Lado Baleato el tiempo que dedicaron a reunirse con nosotros y sentar las bases de este trabajo. También quiero agradecer la labor de Jose durante el desarrollo del mismo y por todo lo que me ha enseñado, conocimientos que sin duda me serán de gran utilidad más allá de este proyecto. También agradecer a todo el equipo de Biostatech que me recibió con los brazos abiertos desde el primer día, por su implicación en este TFM y en mi formación más allá de lo que estaban obligados a hacer. Gracias Julia, Gael (para mi sigues siendo parte del equipo), Jesús, Manuel y Juan por esas conversaciones durante el café. También agradecer a Vicente la oportunidad que me ha dado de seguir formando parte de Biostatech una vez finalizado este TFM. Por supuesto estoy más que agradecido a todos mis amigos, los de Asturias por mantenerme al día estos casi dos años que llevo en Galicia y a los que he hecho aquí que han conseguido darme razones suficientes para no irme a Asturias todas las semanas, que no es poco. Finalmente tengo que agradecer a mi padre todo lo que lleva hecho, hace y hará por mí, incluso aunque te odie un poco más después de cada partida a la escoba. Sí estos agradecimientos fueran ponderados no tendría páginas suficientes para ti. No me faltan motivos y gente por la que estar agradecido cada día y por ello, de nuevo, gracias.

Índice general

Resumen

Resumen en español

El cáncer es una enfermedad responsable de un alto nivel de mortalidad a nivel global. Su detección precoz es clave para la recuperación del paciente ya que, en estadios tardíos de la enfermedad especialmente, los únicos tratamientos son realmente invasivos y su aplicación puede llegar a ser no viable por la propia salud del paciente. Por ello la investigación para desarrollar métodos baratos, no invasivos y eficientes a la hora de detectar potenciales casos de desarrollo de la enfermedad es de especial importancia. En este trabajo se tratará de utilizar la información provista por los perfiles de N-glucanos de pacientes entre otras variables recogidas en el proyecto AEGIS. Para ello se aplicarán técnicas del análisis de datos composicionales. Primero se dará un marco teórico para trabajar en este contexto y se compararan con técnicas del análisis multivariante tradicional. Adicionalmente se utilizan los balances de las distintas partes que conforman la composición. Una vez establecidos distintos clasificadores para detectar pacientes que hayan fallecido a causa de un cáncer, se muestra como el uso de datos composicionales suponen una mejoría de los resultados en multitud de contextos.

English abstract

Cancer is a disease responsible for a high level of mortality globally. Early detection is key to the patient's recovery, as in late stages of the disease, the only available treatments are highly invasive and may not be viable due to the patient's health condition. Therefore, research to develop cheap, non-invasive, and efficient methods for detecting potential cases of cancer development is of utmost importance. This study aims to use the information provided by N-glycan profiles of patients, among other variables collected in the AEGIS project. Compositional data analysis techniques will be applied for this purpose. First, a theoretical framework will be provided to work in this context and compared with traditional multivariate analysis techniques. Additionally, balances of the different parts that make up the composition are used. Once different classifiers are established to detect patients who have died due to cancer, it is demonstrated how the use of compositional data leads to improved results in a multitude of contexts.

Capítulo 1

Introducción

El cáncer representa una de las principales causas de mortalidad a nivel mundial, con un estimado de 10 millones de muertes en 2020 (Ferlay et al., 2021). Asimismo, en España figura como una de las principales causas de defunción, en el año 2022 representó el 25 % de la mortalidad registrada a lo largo del año, llegando a la cifra de 115.000 fallecidos (INE, 2023). Caracterizándose por ser una proliferación de células que han logrado evadir los mecanismos de control endógeno centrales, el cáncer se clasifica según su origen y gravedad. En el último año, el cáncer de mama es el tipo de cáncer más frecuentemente diagnosticado y la principal causa de muerte por cáncer en mujeres, representando el 23 % de todos los casos de cáncer y el 14 % de las muertes relacionadas con esta enfermedad. Por otro lado, el cáncer de pulmón es la principal causa de muerte en hombres, comprendiendo el 17 % de todos los casos nuevos de cáncer y el 23 % de las muertes totales por cáncer. Aunque la letalidad ha descendido en los últimos 30 años gracias a terapias más efectivas y al abordaje temprano, se pronostica que habrá 28 millones de nuevos casos de cáncer en todo el mundo cada año para 2040, si la incidencia se mantiene estable y el crecimiento y el envejecimiento de la población continúan de acuerdo con las tendencias recientes (Ferlay et al., 2021). La capacidad de diagnóstico temprano en el cáncer viene determinada por la utilización de biomarcadores. Un biomarcador tumoral ideal será el que sea fácil de obtener de forma nada (p.e. orina) o poco invasiva (p.e. muestra sanguínea), barato de analizar y que tenga una alta fiabilidad y sensibilidad (Mollarasouli et al., 2022). Si bien esto puede resultar idílico, en los últimos años la glucómica ha emergido como una disciplina destacada en el ámbito de las ciencias biológicas y médicas, ofreciendo la promesa de comprender exhaustivamente, caracterizar y tratar eficazmente las enfermedades humanas (Lauc et al., 2016, O'Flaherty et al., 2022). Concretamente, investigaciones recientes, lograron establecer relaciones entre los N-glucanos plasmáticos y el cáncer (Adamczyk et al., 2012, Gebrehiwot et al., 2018). Muchas de estas nuevas técnicas analíticas obtienen como respuesta la abundancia relativa. Este tipo de datos son catalogados comúnmente como datos composicionales, ya que la señal obtenida integra 1 y que se discretiza para dar los distintos glucanos (Saldova et al., 2014). Aitchison desarrolló la primera definición de datos composicionales, su tratamiento y limitaciones que presentaban a la hora de aplicar las técnicas estadísticas más comunes (Aitchison, 1982). Clásicamente se han propuesto una serie de transformaciones o el uso de balances para solventar estos problemas (Van Den Boogaart y Tolosana-Delgado, 2013). Asimismo, en los últimos años se han desarrollado otras herramientas como por ejemplo nuevos algoritmos basados en el uso de balances(Rivera-Pinto et al., 2018). Por lo tanto, en el presente estudio exploramos las diferentes aproximaciones a datos composicionales con el objetivo de crear un sistema de clasificación supervisado que permita distinguir a los pacientes con un alto riesgo de fallecer de cáncer y los compararemos con métodos de análsis multivariante clásico.

1.1. Motivación

Los datos de los que partimos provienen de cromatografías en las que para cada GP se indica el porcentaje que representa en la muestra, al tratarse de datos de porcentajes, como indicábamos al principio de este texto, la primera definición de datos composicionales encaja a la perfección y ya solo nos indica que hacer uso de las técnicas provenientes del análisis composicional pueden ser útiles. Pero aunque esa primera definición ya sea suficiente justificación para explorar el enfoque composicional tener en cuenta la segunda definición nos ayuda a tener un objetivo de estudio claro desde el principio.

Como se ha establecido previamente la segunda definición hace especial hincapié en las proporciones entre variables. Bajo este punto de vista no nos interesa tanto el ver que un GP concreto llegue a niveles muy altos, sino la relación que hay entre ese mismo GP con los otros, es decir no nos centraremos tanto en si el GP_3 llega a representar el 15 % sino que nos interesaría ver por ejemplo que el porcentaje del GP_3 es el doble que el del GP_6.

Aunque entraremos más en detalle a continuación, uno de los principales problemas con el que se topan los investigadores al tratar de hacer uso de datos composicionales es la presencia de ceros en la muestra. Esto se debe mayormente a que dos de las herramientas más comunes en este campo son los logaritmos y los ratios entre partes, ambas funciones están no definidas en el 0 (en el caso de los ratios cuando es el denominador). Por fortuna nuestros datos son estrictamente positivos y por tanto no tendremos que hacer frente a este problema, lo cuál es una razón menos para no recurrir al análisis composicional.

Capítulo 2

Material y métodos

2.1. Proyecto AEGIS

La fuente de datos del presente estudio se basa en el estudio AEGIS, "A-Estrada Glycation and Inflammation Study" (Gude, 2015). El estudio se realizó en el municipio de A Estrada (42°41' 21"N, 8°29' 14"W, figura 2.1), situado en el noroeste de la península ibérica, con una población censal de 19346 habitantes. Aproximadamente el 75 % de la población viven en el entorno rural. La selección de los participantes ($n_{grupo} = 500$) se realizó mediante un muestreo aleatorio estratificado por décadas de vida del Registro de la Tarjeta Sanitaria, que tiene una cobertura superior al 95 % de la población. Todos los individuos que no pudieron dar consentimiento informado fueron excluidos, la mayoría de ellos por demencia, enfermedades en fase terminal e incapacidad para la comunicación. Finalmente, un total de 1513 sujetos participaron en el estudio [tasa de participación, 68 %; 678 (45 %) hombres y 838 (55 %) mujeres]. La participación fue más baja en hombres (65 %) que en mujeres (71 %), similar a como los jóvenes presentaron una menor tasa de participación que los mayores. No hubo diferencias significativas en edad, sexo y lugar de residencia (rural/urbano) entre los que participaron en el estudio y aquellos que no lo hicieron. Desde noviembre de 2012 hasta marzo de 2015, todos los sujetos fueron contactados sucesivamente y se les pidió que asistieran al Centro de Atención Primaria local para la evaluación (2.1), que incluyó un cuestionario estructurado administrado por el entrevistador y la extracción de una muestra de sangre venosa en ayunas. Complementariamente, los datos de exitus por cáncer de esta cohorte fueron recogidos en septiembre de 2023. Estos datos fueron asociados a los datos de la cohorte y pseudoanonimizados. La mediana de edad de los participantes fue de 52 años (rango, 18-91 años). Las características básicas de hombres y mujeres se resumen en la tabla **??**.

A continuación se muestran las características demográficas que utilizaremos durante el estudio agrupando los pacientes en función de la variable CCM. La tabla **??** muestra las frecuencias absolutas de cada sexo según CCM. Es notable como, a pesar de representar un número total menor en el estudio, los hombres representan la mayoría de pacientes que han sido víctimas mortales de un cáncer. Como se mencionó anteriormente las variables que utilizamos para estimar los distintos modelos de clasificación son GPs, edad, sexo e IMC de cada individuo. En la figura 2.2 se puede apreciar como los fallecidos por cáncer presentan una edad media superior (67.10 años) que el resto de la muestra (52.10 años). Es

Sexo / Variable	Mujer	Hombre
IMC (kg/m^2)	27.33 (23.83-31.38)	28.28 (25.36-31.34)
Edad (Años)	53 (40-67)	52 (38-66.75)

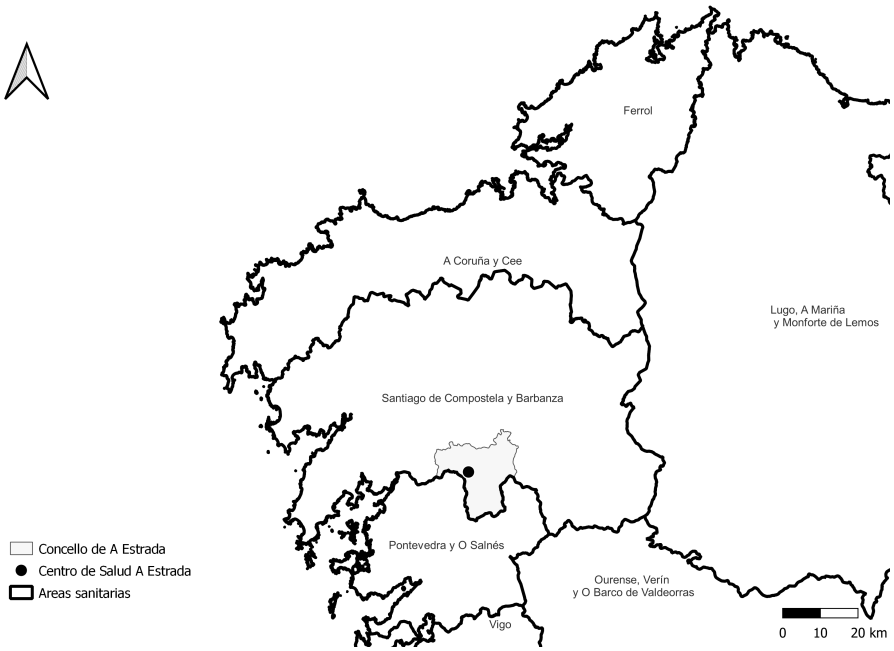


Figura 2.1: Mapa donde se muestra la situación geográfica del concello de A Estrada y su centro de salud

Sexo \ CCM	No	Sí
Mujer	818	17
Hombre	653	25

Cuadro 2.1: Tabla de confusión entre la variable Sexo y la variable CCM (Cáncer Causa de Muerte)

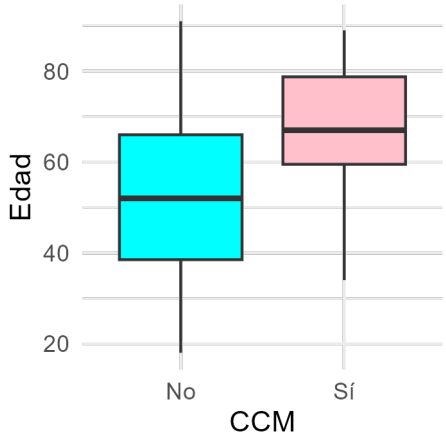


Figura 2.2: Gráfico de cajas de 'Edad' desglosada por CCM (rosa para víctimas de cáncer, azul cian para el resto)

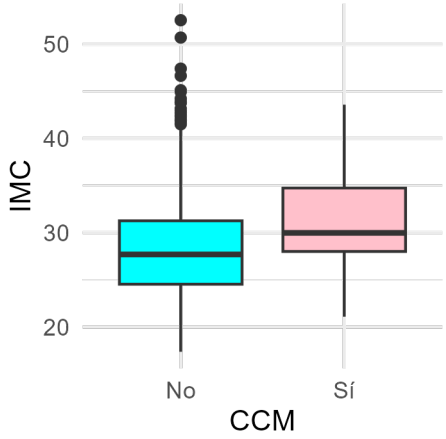


Figura 2.3: Gráfico de cajas de IMC desglosado por CCM (rosa para víctimas de cáncer, azul cian para el resto)

algo esperable pues, como en la mayoría de enfermedades, una edad avanzada supone potencialmente un factor de riesgo importante. En cuanto al IMC, el valor mínimo registrado fue 17.36 y el máximo de 52.54, la media global se sitúa en 28.24 y esta variable tiene una desviación típica de 11.88. En la figura 2.3 tenemos un gráfico de cajas de esta variable agrupando los individuos según CCM. A simple vista parece que las personas fallecidas por cáncer presentaban un tendencia a valores altos del IMC, esta discrepancia se ve reflejada en las medias de cada grupo, siendo de 28.16 en el grupo control y de 31.13 en grupo de fallecidos por cáncer.

La extracción de sangre en todos los pacientes se realizó después de una noche de ayuno utilizando BD Vacutainer®SST Serum Separation Tubos (BD, Plymouth, Reino Unido). Después de la recolección, los tubos de muestra se invirtieron cinco veces, se dejaron 30 minutos para la coagulación y se centrifugaron a temperatura ambiente durante 15 min a 1300 × g en un ángulo fijo centrífugo. El suero se extrajo, se dividió en alícuotas y se almacenó a -80 ºC para su uso posterior.

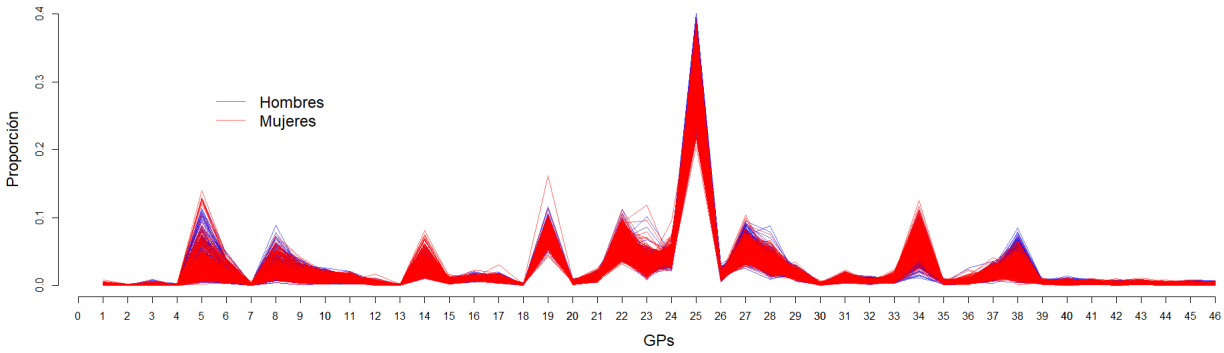


Figura 2.4: GPs desglosados por sexo

Los N-glucanos se liberaron a partir de 5 µl de muestras de suero utilizando un método automatizado de alto rendimiento modificado (Stöckmann et al., 2015). Para la cromatografía de interacción hidrofílica (HILIC) y la cromatografía líquida de ultra alta resolución (UPLC), los N-glucanos marcados con fluorescencia se separaron en un instrumento Waters Acquity H-Class UPLC de Waters (Waters, Milford, MA, EE.UU.). El instrumento estaba bajo el control del software Empower 3, build 3471 (Waters, Milford, MA, EE.UU.). Los N-glucanos marcados se separaron en una columna de cromatografía de glucanos BEH Ethylene Bridged Hybrid de Waters, 150 × 2,1 mm i.d., partículas BEH de 1,7 µm (Waters, Milford, MA, EE.UU.). El método de separación utilizó un gradiente lineal. Las muestras se mantuvieron a 4 ºC antes de la inyección. El sistema se calibró utilizando un patrón externo de oligómeros de glucosa hidrolizados y marcados con 2-AB para crear una escalera de dextrano, como se ha descrito anteriormente (Royle et al., 2008). Se ajustó una curva de distribución polinómica de quinto orden a la escalera de dextrano para asignar valores de unidades de glucosa (GU) a partir de los tiempos de retención (utilizando el software Empower de Waters. Milford, MA, EE.UU.). Todos los cromatogramas se separaron de la misma manera en 46 picos de acuerdo con (Saldova et al., 2014), y la cantidad de glucanos en cada pico fue expresada como % del área total integrada. Las estructuras de los glucanos se anotaron utilizando la nomenclatura SNFG y el software DrawGlycan-SNFG (University at Buffalo, Buffalo, NY, USA) (Cheng et al., 2017, Neelamegham et al., 2019).

Para representar los GPs optamos por representarlos como si fueran datos funcionales. Para ello hicimos uso del paquete *fda.usc* (Febrero-Bande y Oviedo de la Fuente, 2012). En el eje X se representan cada uno de los GPs y en el eje y la abundancia relativa que constituyen en la muestra. Pese a que se haya utilizado esta representación, no hay que olvidar que en este trabajo estamos tratando de utilizar un enfoque composicional y por ello, cuando más adelante se hable de la media de los GPs estaremos hablando de la media composicional descrita en la ecuación 2.1 y no de la media habitual del análisis de datos funcionales. En la figura 2.4 se representan los GPs de todos los pacientes diferenciando por sexo. En esta gráfica se pueden apreciar dos puntos importantes. Por un lado la estructura general de estos datos ya que parece que todos los individuos siguen un patrón relativamente similar, y por otro que la visualización de datos de dimensión tan alta deja de ser tan intuitiva como en dimensiones bajas. Además cabe señalar que no parecen apreciarse diferencias notables entre un grupo y otro, pero esto podría deberse al exceso de información en la gráfica. Para tratar de condensar esta información se presenta en la gráfica 2.6 la media (composicional) de cada uno de los grupos, de nuevo no parecen apreciarse diferencias notables, como mucho una ligera diferencia a favor de las mujeres en GP34. Haciendo lo análogo pero separando los grupos según su causa de mortalidad se presentan los datos en la figura 2.7. No parece que haya presentes diferencias muy claras, aunque podemos encontrar pequeñas discrepancias en GP5, GP14 y GP22.

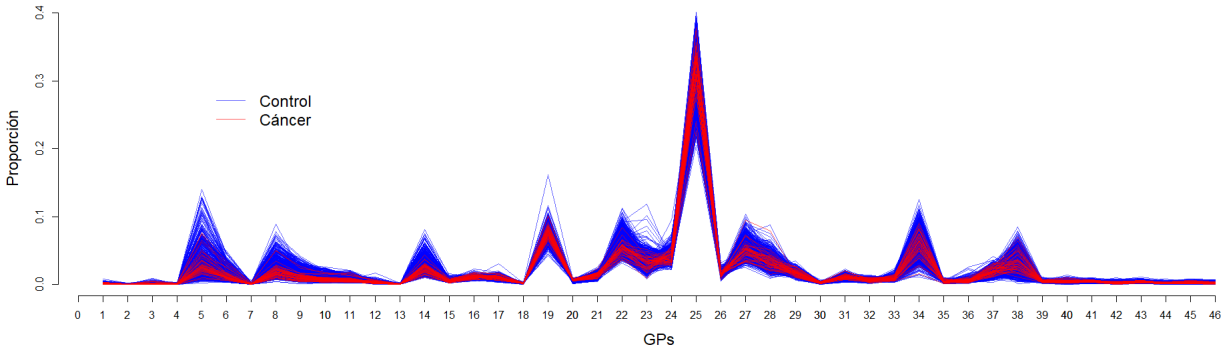


Figura 2.5: GPs desglosados por causa de mortalidad

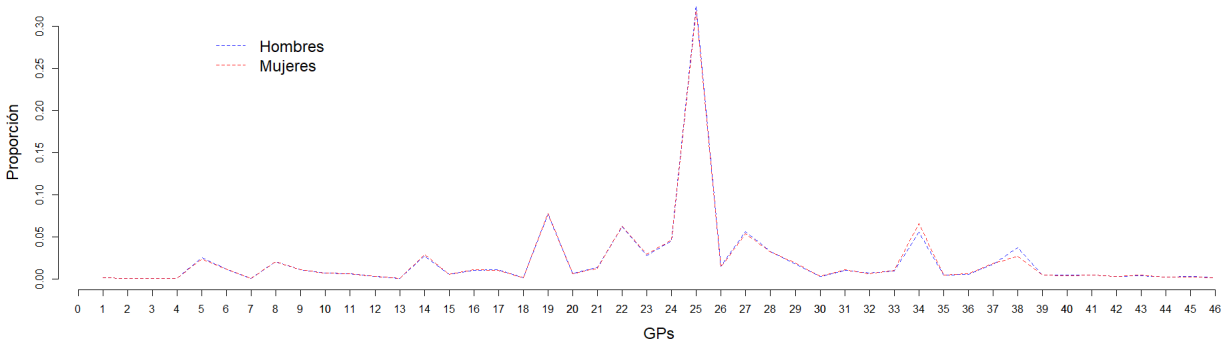


Figura 2.6: Media composicional de GPs en hombres y mujeres

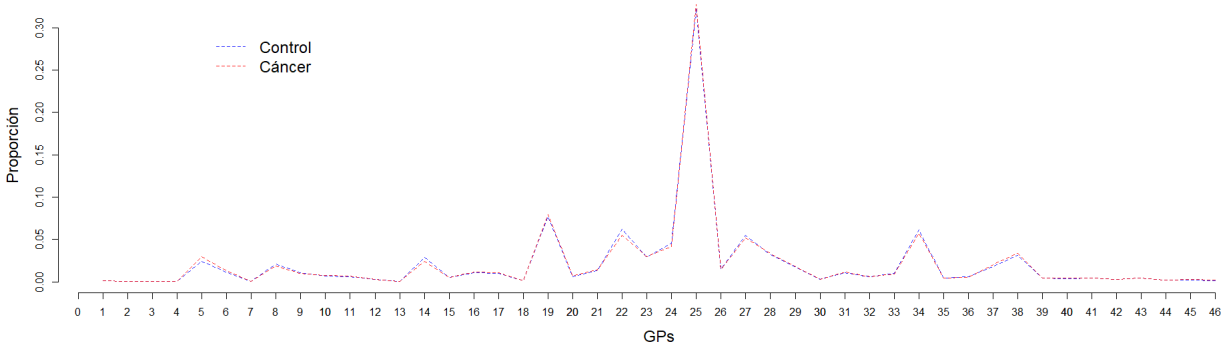


Figura 2.7: Media composicional de GPs en fallecidos por cáncer y grupo control

2.2. Datos composicionales

2.2.1. Motivación de uso

Los datos composicionales se pueden definir en una primera instancia como aquellos datos multivariantes no negativos cuya suma está fijada en alguna constante para todos los individuos. Los ejemplo más habituales en los que se da esta situación surgen a la hora de trabajar con proporciones o porcentajes. En este último caso tendríamos para cada individuo una serie de variables, cada una de ellas con un porcentaje asociado, de manera que al sumar los porcentajes de todas las variables de un mismo individuo obtendríamos un 100 %, en el caso de trabajar con proporciones sería una situación análoga pero con suma constante igual a 1. Esta primera definición es suficiente para la mayoría de situaciones, sin ir más lejos que en nuestro caso con los N-glucanos, donde la señal de cada individuo integra a 1, es decir, tiene una suma constante. Sin embargo surge una segunda definición más general para superar alguna de las limitaciones que implica limitarse a la anterior y es que, puede ocurrir algo tan sencillo como que los datos que hayamos obtenido hayan sido recolectados de tal manera que no recogen información en porcentajes ni proporciones, sino que se componen de medidas absolutas. En esta situación podríamos pasar fácilmente a proporciones dividiendo cada individuo entre el total de la suma de sus variables, pero de esta manera estaríamos perdiendo algo de información. Según este nuevo punto de vista, los datos composicionales se definen como aquellos datos multivariantes no negativos donde la información relativa entre las distintas variables es de especial interés. De esta manera no solo estamos abriendo la posibilidad de trabajar con datos composicionales en casos donde la definición anterior no se cumpliría, sino que además estamos poniendo de manifiesto que el enfoque que daremos será el de estudiar la relación (entendida como proporción) entre los valores de distintas variables. Esta segunda definición , más general, pone en el foco de atención la perspectiva con la que se llevará a cabo el análisis antes que la propia naturaleza de los datos.

Para tratar este tipo de datos es tentador recurrir a las metodologías habituales de análisis multivariantes. Sin embargo, procediendo de esta manera surgen gran variedad de problemas, algunos más sofisticados que otros y que por tanto requieren de cierto cuidado para poder ser resueltos. En la literatura (Pawlowsky-Glahn et al., 2015) se ilustran numerosos ejemplos para la gran variedad de posibles problemáticas. De entre todos estos inconvenientes juzgamos especialmente interesantes los siguientes.

La escala de las proporciones

Supongamos que tenemos datos sobre el porcentaje de jóvenes que cursan estudios universitarios en dos países distintos y que hacemos un seguimiento a lo largo de los años de dichas cifras. Si planteamos que en el primer año tenemos que en el país A el 60 % de los jóvenes ingresan en la universidad mientras que en el país B solo un 1 % , y que en el segundo año en A pasamos a un 62 % y que en B pasamos a un 3 %, el incremento en ambos países "solo"ha sido del 2 %. A pesar de que el porcentaje se ha visto incrementado en ambos países de igual manera, es claro que para el país B el incremento es mucho más significativo que para el país A. Además si nos centrásemos en estudiar el porcentaje de jóvenes que *no* entran en el sistema universitario deberíamos llegar a las mismas conclusiones ya que en el fondo representan la misma información (sabiendo el porcentaje de alumnos que no cursan estudios universitarios sabríamos también el porcentaje que sí los cursa). Pero si nos centramos en esta contraparte negativa tendríamos que en el país A pasamos de un 40 % a un 37 % mientras que en país B pasaríamos de un 99 % a un 97 %. En este caso el cambio más significativo parece que se ha dado en el primer país pero no hay tanta diferencia como podíamos interpretar en el primer planteamiento. Es decir, partiendo de los mismos datos y basándonos en un mismo criterio (comparar porcentajes de manera relativa) hemos llegado a dos conclusiones notablemente distintas. Uno de los objetivos de los datos composicionales será el tratar de evitar este tipo de incongruencias.

Intervalos de confianza fuera de soporte

	Vacunados	**No vacunados**
Jóvenes	46/52 (88.46 %)	311/359 (86.63 %)
Adultos	321/403 (79.65 %)	72/96 (75 %)
Total	367/455 (80.66 %)	383/455 (84.18 %)

Cuadro 2.2: Ejemplo paradoja de Simpson

Si tenemos datos que por ejemplo midan el porcentaje de azúcar de distintas bebidas. Si tuviésemos porcentajes de azúcar en distintas bebidas que fuesen por ejemplo 0, 0, 4, 5, 6, 10, 12, 12 tendríamos que la media es 6.125 y la desviación típica 4.853. De esta manera un 2s-intervalo para le media sería [-3.581,15.831] lo cuál no tiene mucho sentido porque sabemos que es imposible que ninguna proporción esté por debajo de 0 y mucho menos la media. Podríamos truncar el intervalo a [0,15.831], de hecho es una práctica habitual, pero en ese caso no quedaría claro como proceder con la probabilidad que se asigno a la parte del intervalo truncada. Además, en otros casos los intervalos de confianza surgen de asumir alguna distribución particular, si dicha distribución sugiere valores que sabemos que no son posibles, cabe cuestionar la fiabilidad de los resultados basados en asumir dicha distribución como cierta.

Paradoja de Simpson

No es un problema exclusivo de trabajar con datos composicionales pero al trabajar con proporciones es normal que aparezca esta conocida problemática. Para ilustrarla supongamos que tenemos una vacuna para cierta enfermedad y que estudiamos los efectos de dicha vacuna en jóvenes y adultos, contabilizando la cantidad de individuos que superaron la enfermedad sin ningún tipo de complicaciones para cada uno de los cruces entre las variables edad (jóvenes/adultos) y tratamiento (vacunado/no vacunado)

En la tabla **??** podemos ver como, analizando por separado los dos grupos de edad, es claro que el hecho de ser vacunado supone un impacto positivo, ya que consigue un mayor porcentaje de pacientes que superan la enfermedad sin complicaciones. Sin embargo, al estudiar los grupos de edad conjuntamente, parece que el hecho de no ser vacunado repercute en un mayor porcentaje de pacientes que superan la enfermedad sin problema. La explicación matemática que hay detrás de esta aparente paradoja es que las fracciones que definen los porcentajes en cada grupo tienen distintos denominadores y por tanto no se sumarían de manera directa, es decir, si denotamos por $s_{jv}, s_{jn}, s_{av}, s_{an}$ los pacientes sanos con los índices indicando si son de un grupo u otro (j- jóvenes, a-adultos, v-vacunados, n-no vacunados) y por $N_{jv}, N_{jn}, N_{av}, N_{an}$ los pacientes totales en cada grupo siguiendo la misma norma de etiquetación, cuando comparamos por grupos de edad tenemos $\frac{s_{jv}}{N_{jv}} > \frac{s_{jn}}{N_{jn}}$ y que $\frac{s_{av}}{N_{av}} > \frac{s_{an}}{N_{an}}$ pero no ello no implica que $\frac{s_{jv}+s_{av}}{N_{jv}+N_{av}} > \frac{s_{jn}+s_{an}}{N_{jn}+N_{an}}$ ya que $\frac{s_{jv}+s_{av}}{N_{jv}+N_{av}} \neq \frac{s_{jv}}{N_{jv}} + \frac{s_{av}}{N_{av}}$ y $\frac{s_{jn}+s_{an}}{N_{jn}+N_{an}} \neq \frac{s_{jn}}{N_{jn}} + \frac{s_{an}}{N_{an}}$

En esencia, este problema tan común se debe básicamente a aplicar la suma habitual sobre porcentajes ignorando que realmente un porcentaje es una fracción. Esta junto a otras razones que iremos viendo motivan la necesidad de construir un marco teórico matemático adecuado para este tipo de situaciones. Pero antes de entrar en más detalle sigamos viendo algunos de los ejemplos de problemáticas más comunes.

Correlación Espuria

Uno de los problemas que originalmente puso de manifiesto la necesidad de un nuevo marco teórico para tratar con este tipo de datos. La primera mención a este problema data de finales del siglo XIX

Cuadro 2.3: Correlaciones entre las variable originales

	X	Y	Z
X	1	1	0
Y	1	1	0
Z	0	0	1

Cuadro 2.4: Correlaciones entre las variables vistas como partes de una composición

	X	Y	Z
X	1	1	-1
Y	1	1	-1
Z	-1	-1	1

(Pearson, 1897). Para ilustrar a qué se refería Pearson en ese artículo basta un ejemplo muy sencillo. Supongamos que tenemos 3 variables, donde la primera es exactamente el doble de la segunda y la tercera es completamente independiente del resto. Si calculásemos la matriz de correlaciones teórica sería:

Sin embargo, si nosotros quisiésemos trabajar con las proporciones, es decir, tratando cada variable como una parte de una composición total, podríamos reescalar los valores a 1. Esto provocaría que, si calculásemos las correlaciones como hicimos cuando trabajábamos con las variables originales, obtendríamos una matriz 2.4con cambios importantes respecto a la matriz 2.3.

La diferencia es clara, las correlaciones de Z con el resto de variables pasan de 0 a -1, pasamos de independencia (lineal) total al extremo opuesto, dependencia lineal total. Esto se debe al reescalado que hemos realizado, forzando que la suma de las tres variables sea siempre 1 tenemos que $X + Y + Z = 1$ pero como habíamos establecido antes $Y = 2X$ y por tanto, un poco de álgebra básica establece que $Z = 1 - 3X = 1 - \frac{3}{2}Y$. Claro, uno podría pensar que realmente el problema ha surgido por aplicar el reescalado, sería tan sencillo como no aplicarlo y así evitar este problema. Hay dos motivos para no huir de esta manera. El primero es que hay casos en los que ya los datos de los que partimos vienen expresados en forma composicional, por tanto ese problema de que las correlaciones se ven afectadas espuriamente por la restricción de tener una suma constante no es algo que hayamos creado nosotros al tratar los datos y por tanto no nos queda más remedio que aceptarlo. Por otro lado, en los casos en los que sí podría estar en nuestra mano no aplicar los datos composicionales, estaríamos ignorando un amplio abanico de opciones de cara al resto de análisis estadístico que abre el hecho de trabajar con datos composicionales, opciones entre las cuáles se incluyen herramientas para tratar con estas correlaciones espurias.

Este último problema es de especial interés pues el estudio de los datos composicionales se inicia cuando Karl Pearson hizo públicas las correlaciones espurias que encontró al utilizar herramientas estadísticas estándar con datos de proporciones. Esta publicación data de 1897 pero no fue hasta 1960

cuando se empezó a advertir del peligro de utilizar metodologías estándar multivariantes con datos composicionales (Chayes, 1960). En estos momentos el problema que hemos visto de la clausura se veía simplemente como un factor que introducía un sesgo negativo en las correlaciones. El objetivo por tanto, era tratar de separar la correlación *real* entre las variables de esta correlación espuria debida a la restricción de suma constante. Destacan en este punto del desarrollo de datos composicionales varios artículos (Sarmanov y Vistelius, 1959) en un contexto geológico y (Mosimann, 1962) en un contexto biológico.

Hasta este punto del desarrollo del análisis de datos composicionales, el objetivo de estudio normalmente era la distorsión que se producía en los resultados al emplear herramientas típicas del análisis multivariante, el punto de inflexión se produjo cuando Aitchison (Aitchison, 1982) vio los datos composicionales desde un nuevo enfoque. Él proponía centrarse en que los datos composicionales daban información relativa entre las distintas partes que conforman una composición. Para ello empleó los log-ratios, ya que por un lado los ratios son idóneos para estudiar la relación entre las distintas partes y por otro, el uso de los logaritmos a los ratios resultan en una mayor operabilidad que con los ratios a solas. Utilizando los log-ratios, propuso varias transformaciones que conseguían llevar los datos composicionales del espacio símplex, donde se encuentran originalmente debido a la restricciones de no negatividad y suma constante, al espacio real multivariante. De esta manera se abría la puerta a aplicar la artillería del análisis multivariante. Más adelante desarrollaremos algunas de estas transformaciones en detalle planteando sus ventajas y desventajas.

Es interesante ver como realmente la idea no es tan única como pudiera parecer, para cualquiera que haya llevado a cabo una regresión logística no es ninguna novedad transformar los datos mediante alguna función que permite trabajarlos con mayor comodidad, pudiendo deshacer el camino de dicha transformación para una mayor interpretabilidad de los resultados una vez obtenidos.

La última fase del desarrollo de los datos composicionales, en la que nos encontramos actualmente, se inicia al comienzo de este siglo cuando, de manera independiente, se formalizó la estructura algebraica del símplex justificando, no solo que tiene estructura de espacio vectorial, sino que llega a ser incluso espacio de Hilbert en dos artículos distintos (Billheimer et al., 2001,Pawlowsky-Glahn y Egozcue, 2001). La idea en este punto era utilizar esta estructura matemática para desarrollar herramientas que se pudiesen aplicar dentro del contexto del símplex, sin necesidad de aplicar una transformación que llevase los datos a un espacio real. Pese a ello esta manera de actuar sigue siendo muy similar a las transformaciones que propuso originalmente Aitchison, pues muchas veces lo que se consigue trabajando dentro del símplex es encontrar una base de este espacio, y una vez establecida esta base, los datos pueden expresarse en función de ella. Esta nueva representación de los datos consistirá en representar cada dato como una serie de coordenadas, estas coordenadas ya no estarán bajo la restricción de suma constante, por tanto son susceptibles de ser tratadas como datos multivariantes, como ocurría cuando Aitchison aplicaba sus transformaciones más directas.

2.2.2. Principios de invarianza

Una vez visto de manera resumida el desarrollo histórico de los datos composicionales y algunos ejemplos en los que el uso de los mismos puede ser más que adecuado necesitamos establecer unas ciertas normas que deben cumplir para mantener cierto rigor en un análisis estadísticos que recurra a los datos composicionales. Estas normas se constituyen de tres principios de invarianza. Los dos primeros no suelen suponer ningún problema, sin embargo el tercero veremos que tiene implicaciones algo más severas.

Principio de invarianza por permutación
Este primer principio se resume en que el orden de las variables no debería alterar los resultados

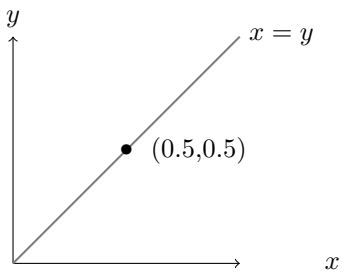


Figura 2.8: El punto (0,5, 0,5) es el representante de la clase formada por todos las composiciones de dos partes que tienen sus dos partes iguales, representados en la semirrecta x=y que empieza en (0,0)

de un análisis estadístico, esto es, la composición $\mathbf{x} = (x^1, x^2, ..., x^D)$ contiene la misma información que la composición $\mathbf{x}* = (x^2, x^1, ..., x^D)$ y que cualquier otra reordenación de sus componentes.

Principio de invarianza por escala
Este principio indica que las composiciones mantienen la misma información independientemente del reescalado que se le aplique. Esto es, $\mathbf{x} = (x^1, x^2, ..., x^D)$ contiene la misma información que $\mathbf{x}* = (k * x^1, k * x^2, ..., k * x^D)$ siendo k un número real estrictamente positivo. Una manera más matemática de entender esta idea es que una composición se encuentra dentro de una clase de composiciones. Y nosotros lo que haremos sera estudiar los representantes de cada clase, siendo el representante de cada clase una composición cuya suma se haya fijado previamente, por ejemplo 1 si se prefiere trabajar en proporciones. El resto de elementos de esa clase serían reescalados de este representante. En ?? se muestra un ejemplo de esta representación de las clases de manera visual.

Coherencia subcomposicional Para muchas técnicas estadísticas es imprescindible tener definida una distancia entre los individuos que se quieran estudiar, métodos de *clustering*, de clasificación, de detección de valores atípicos... Tener una distancia nos permite cuantificar una noción tan básica como es saber cómo de lejos está una entidad de otra, un libro de una estantería, una persona de un grupo de perros, las previsiones de ventas de una empresa en un año con las ventas reales de dicho año. Para el caso de análisis multivariante hay definidas una basta cantidad de ellas con estudios pormenorizados de sus ventajas y desventajas en la literatura. Es importante detenerse a considerar este aspecto ya que como hemos establecido anteriormente mucho métodos dependen de la distancia utilizada.

Debida a la semejanza de los datos composicionales con los datos multivariantes uno podría verse tentado a utilizar cualquiera de estas distancias, veremos que no sería una manera adecuada de proceder en la mayoría de casos. Pues a la hora de elegir una distancia con la que trabajar no solo debemos asegurarnos de que cumpla la definición matemática de distancia, sino de interpretarla, la distancia discreta y la euclídea son distancias de igual validez, pero sus diferencias son más que notables y no pueden usarse e interpretarse indistintamente.

El problema se presenta cuando tratamos con subcomposiciones. Una subcomposición se define como una composición a la que se le suprime una o más partes. Por ejemplo si $\mathbf{x} = (x^1, x^2, x^3)$ es una composición, $\hat{\mathbf{x}} = (x^1, x^2)$ podría ser una subcomposición suya.

Una manera intuitiva de entender la problemática es la siguiente. Cuando utilizamos distancias para comparar individuos nos interesa obtener distancias grandes para individuos muy dispares y distancias pequeñas cuando los individuos son más similares entre sí. De esta manera, si tenemos dos composiciones $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$, una distancia $d(\mathbf{x}, \mathbf{y})$ para compararlas, $\mathbf{x}^*, \mathbf{y}^*, \in \mathcal{S}^P$ con $P < D$ sus dos subcomposiciones correspondientes y $d^*(\mathbf{x}^*, \mathbf{y}^*)$ la distancia d ajustada a la nueva dimensión, de-

bería cumplirse que $d(\mathbf{x},\mathbf{y}) \geq d(\mathbf{x}^*,\mathbf{y}^*)$ es decir, reduciendo la dimensión de los datos, estudiando menos características, deberíamos encontrar como mucho tanta diferencia como estudiando todas las características. Esta sencilla idea, combinada con el principio de invarianza por escala, tiene como consecuencia que una distancia tan prolífera como lo es la distancia euclídea ya no sea adecuada. Veamos un ejemplo:

Tenemos los individuos $\mathbf{x} = (0{,}5, 0{,}2, 0{,}3)$ $\mathbf{y} = (0{,}5, 0{,}3, 0{,}2)$. La distancia euclídea entre ellos sería $d(\mathbf{x},\mathbf{y}) = 0{,}141$. Ahora bien, si nos quedamos con las subcomposiciones resultantes de suprimir la primera componente $\mathbf{x}^* = (0{,}2, 0{,}3)$, $\mathbf{y}^* = (0{,}3, 0{,}2)$. Haciendo uso del principio de invarianza podríamos reescalar estas proyecciones para trabajar con proporciones, obteniendo para cada proyección $\mathbf{rx}^* = (0{,}4, 0{,}6)$, $\mathbf{ry}^* = (0{,}6, 0{,}4)$. Pero la distancia euclídea entre ambos ahora sería $d(rx^*, ry^*) = 0{,}283 > d(\mathbf{x},\mathbf{y}) = 0{,}141$.

Esto se puede interpretar como que estudiando menos características entre los individuos podemos llegar a encontrar más distancia. Esto es especialmente peligroso a la hora de aplicar técnicas de reducción de dimensión ya que podríamos llegar a quedarnos con muchas menos componentes de las que realmente harían falta para tener una imagen fiel, aunque simplificada, de los datos.

2.2.3. Geometría Aitchison

Como hemos establecido al principio, el enfoque de los datos composicionales fundamentalmente se centra en el estudio de las relaciones entre componentes. Esta idea junto a los principios y los inconvenientes que aparecen al trabajar con distancias habituales, crea la necesidad de establecer un marco teórico que proporcione las herramientas matemáticas necesarias para poder analizar nuestros datos. Con este objetivo se plantea la Geometría de Aitchison, (Pawlowsky-Glahn y Egozcue, 2001). Para justificar las propiedades de este espacio debemos empezar explicando cuales son los elementos que lo conforman y las operaciones fundamentales que nos permitirán manipularlo.

Espacio muestral:
El espacio muestral, es decir los posibles elementos con los que trabajaremos quedan confinados en el D-símplex definido por

$$\mathcal{S}^D = \left\{ \mathbf{x} = [x^1, x^2, ..., x*D] | x^i > 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x^i = k \right\}$$

Recordemos que es posible que nuestros datos originales no tengan una suma constante en todos los individuos, veremos que no supondrá ningún problema pues podemos establecer un k arbitrario (puede ser 1 si queremos trabajar con proporciones o 100 si preferimos trabajar en porcentajes, pero puede ser cualquier constante positiva) y aplicando la operación clausura para ese k en todos los individuos conseguiremos que tengan suma constante de manera que queden todos en el mismo símplex.

Clausura:
Dados $\mathbf{x}$ una composición de D partes, es decir, $\mathbf{x} = [x^1, x^2, ..., x^D] \in \mathbb{R}^D$ y $k > 0$ se define la clausura de $\mathbf{x}$ para k como

$$C_k(\mathbf{x}) = \left[\frac{k \cdot x^1}{\sum_{i=1}^{D} x^i}, \frac{k \cdot x^2}{\sum_{i=1}^{D} x^i}, ..., \frac{k \cdot x^D}{\sum_{i=1}^{D} x^i} \right] \in \mathcal{S}^D$$

Con esta operación reescalamos las composiciones de manera que la suma de todas sus partes sea k, con esta operación tenemos una manera de formalizar el concepto que se mencionó anteriormente de clases, haciendo que dos composiciones sean de la misma clase si $C_k(\mathbf{x}) = C_k(\mathbf{y})$ independientemente

de la constante k. En adelante la operación clausura se denotará simplemente por $C()$, sin subíndice, salvo que sea necesario en el contexto.

Amalgamación
Dadas una composición $\mathbf{x} \in \mathcal{S}^D$ y una colección de índices $A = i_1, ..., i_a$ con $D-a \geq 1$, se denomina parte amalgamada a $x_A = \sum_{j \in A} x^j$. A la composición $\mathbf{x}' = [\mathbf{x}_{\bar{A}}, x_A] \in \mathcal{S}^{D-a+1}$ se le denomina composición amalgamada donde $\mathbf{x}_{\bar{A}}$ contiene todas las partas cuyos índices no están en A. Esta operación permite reducir la dimensión de una composición manteniendo más información que si se trabajase con una subcomposición.

Para dotar al símplex de estructura de espacio vectorial y así poder aplicar multitud de métodos que dan esta condición por sentada necesitamos definir operación que hagan el papel de la suma y el producto por escalares.

Perturbación
Dadas dos composiciones $\mathbf{x}$,$\mathbf{y} \in \mathcal{S}^D$ la perturbación de $\mathbf{x}$ por $\mathbf{y}$ se define como

$$\mathbf{x} \oplus \mathbf{y} = C[x^1 y^1, x^2 y^2, ..., x^D y^D]$$

Esta operación jugará el papel de la suma "habitual". Es obvio que la perturbación es conmutativa.

Elemento neutro
El elemento neutro de la operación perturbación, $\mathbb{1} \in \mathcal{S}^D$ se define como aquel que cumple:

$$\mathbf{x} \oplus \mathbb{1} = \mathbf{x} \; \forall \; \mathbf{x} \in \mathcal{S}^D$$

Claramente si tomamos $\mathbb{1}^* = [1, ..., 1]$ se cumple la condición anterior ya que

$$\mathbf{x} \oplus \mathbb{1} = C[x^1 1, ..., x^D 1] = C[x^1, ..., x^D] = \mathbf{x}$$

Notar que $C(\mathbf{x})$ coincide con $\mathbf{x}$ porque $\mathbf{x} \in \mathcal{S}^D$. Sin embargo no tiene por qué cumplirse que $\mathbb{1} \in \mathcal{S}^D$ por eso el elemento neutro será el representanta de la clase de $\mathbb{1}^*$ en el símplex, es decir $\mathbb{1} = C(\mathbb{1}^*)$

Elemento opuesto
Dado $\mathbf{x} \in \mathcal{S}^D$ su elemento opuesto $-\mathbf{x} \in \mathcal{S}^D$ se define como aquel que cumpla:

$$\mathbf{x} \oplus -\mathbf{x} = \mathbb{1}$$

De esta manera se puede utilizar la siguiente notación para denotar algo análogo a la resta habitual de multivariante. Sean $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^D$ entonces:

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus -\mathbf{y}$$

Potenciación
Dados $\mathbf{x} \in \mathcal{S}^D$ una composición y $\alpha \in \mathbb{R}$ una constante, la potenciación de $\mathbf{x}$ por α se define como

$$\alpha \odot \mathbf{x} = C[(x^1)^\alpha, (x^2)^\alpha, ..., (x^D)^\alpha]$$

Se cumple además que el elemento opuesto de $\mathbf{x}$ que denotábamos por $-\mathbf{x}$ se puede calcular como $-\mathbf{x} = (-1) \odot \mathbf{x}$

Esta operación juega el papel del producto por escalares. Se cumple que es distributiva respecto a la perturbación, permitiendo así una traducción directa entre la suma y producto por escalares habituales de multivariante y la perturbación y potenciación. Esto permite que podamos encontrar análogos para

muchas de las medidas que se utilizan habitualmente en análisis multivariante. Una de las más básicas es sin duda la media. Si tenemos $\mathbf{x_1}, ..., \mathbf{x_N} \in \mathcal{S}^D$ composiciones, su media se define como:

$$\bar{\mathbf{x}} = \bigoplus_{i=1}^{N} \frac{1}{N} \odot \mathbf{x_i} \tag{2.1}$$

Cuya estructura, salta a la vista, es muy similar al cálculo de la media en datos multivariantes. De manera similar se pueden encontrar análogos composicionales para otras medidas del análisis multivariante como la varianza por ejemplo. Con lo visto hasta ahora la geometría de Aitchison nos permite trabajar en un espacio vectorial en el contexto de datos composicionales. Pero como mencionamos anteriormente esta estructura llega a ser espacio de Hilbert, lo cuál nos proporciona una herramienta matemática para hablar de independencia entre datos composicionales lo cuál, por la propia naturaleza de los mismos, es un tema delicado, no hay que olvidar nunca que la restricción de suma constante fuerza a que el incremento de una parte repercuta en la disminución de otras inevitablemente. Dadas dos composiciones $\mathbf{xy} \in \mathcal{S}^D$ El producto interior de este espacio que lo dota de estructura de espacio de Hilbert se define de la siguiente manera:

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{D} \sum_{i<j} ln\frac{x^i}{x^j} ln\frac{y^i}{y^j} = \sum_{i=i}^{D} ln\frac{x^i}{g(\mathbf{x})} ln\frac{y^i}{g(\mathbf{y})}$$

Donde $g_m(\mathbf{x}) = (\prod_{i=1}^{D} x_i)^{\frac{1}{D}}$. Esto permite definir independencia entre dos composiciones $\mathbf{x}$, $\mathbf{y}$ si $\langle \mathbf{x}, \mathbf{y} \rangle_a = 0$. Además al tener un un producto interior quedan definidas de manera implícita una norma y una distancia:

$$||\mathbf{x}||_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a \qquad\qquad d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a$$

respectivamente. En (Egozcue et al., 2003) se puede encontrar una explicación más detalla de la estructura algebraica de este espacio así como equivalencias entre estructuras en el el espacio euclídeo, $\mathbb{R}^{D-1}$ y $\mathcal{S}^D$.

2.2.4. Trabajar en coordenadas

Aunque en este texto hemos decidido explicar primero la justificación de la estructura de espacio de Hilbert, esta estructura matemática es de las partes más recientes del desarrollo del análisis de datos composicionales. Como se menciono en la sección 2.2.1, la idea de aplicar transformaciones sobre los datos composicionales para trasladar un problema al análisis multivariante llegó tiempo antes de la mano de Aitchison (Aitchison, 1982). La idea básica para tratar con datos composicionales mediante estas transformaciones es definir una base que realice la transformación de un espacio a otro, aplicar dicha transformación y finalmente trabajar sobre las coordenadas en el nuevo espacio. En función de los propios datos y de los objetivos que se pretendan perseguir será más apropiado una u otra base para llevar a cabo esta transformación. A continuación presentamos algunas de las transformaciones más habituales, entre ellas se encuentran las que emplearemos más adelante para llevara cabo nuestro análisis estadístico.

Log-ratios centrados

CLR, por sus siglas en inglés, es una transformación que, siguiendo la notación ya establecida, actúa de la siguiente forma sobre los datos originales:

$$clr(\mathbf{x}) = \left[ln\frac{x_1}{g_m(\mathbf{x})}, ln\frac{x_2}{g_m(\mathbf{x})}, ..., ln\frac{x_D}{g_m(\mathbf{x})} \right] \tag{2.2}$$

Esta transformación es de las más básicas y mantiene una correspondencia 1-1 entre las componentes originales y las variables tras la transformación, es decir, si partimos de composiciones con D

componentes el nuevo espacio será el espacio euclídeo de dimensión D también. Tener esta correspondencia puede ayudar en la interpretación de los resultados, sin embargo mantiene la restricción de suma constante ya que fuerza que las nuevas coordenadas de cada individuo sumen 0 ya que si denotamos por x_i^* a cada una de las partes del vector $clr(\mathbf{x})$ se cumple que

$$\sum_{i=1}^{D} x_i^* = \sum_{i=1}^{D} ln\frac{x_i}{g_m(\mathbf{x})} = ln\frac{\prod_{i=1}^{D} x_i}{g_m^D} = 0 \tag{2.3}$$

Log-ratios isométricos Denotada por ILR por sus siglas en inglés, esta transformación elimina la restricción de la colinealidad presente entre las componentes originales ya que se libra de toda condición de suma constante. Al mismo tiempo mantiene las distancias entre los individuos, lo cuál es una ventaja a la hora de trabajar con herramientas que tengan en cuenta la estructura espacial de los datos como pueden ser métodos de clustering. Sin embargo, el nuevo espacio en el que se encuentran los datos una vez transformados por esta base tiene una componente menos, perdiendo así la interpretabilidad que tenía la transformación CLR. Pese a ello, este método es el más versátil en la mayoría de ocasiones y es el que se emplea por defecto por los pocos inconvenientes que presenta en la computación de variedad de métodos.

Las coordenadas en esta base no se calculan de una manera tan directa como el caso anterior ya que realmente no hay una única transformación ILR. como sí ocurre con la transformación CLR. Cuando se habla de la transformación ILR realmente se está hablando de una familia de transformaciones, las cuales precisan previamente de una base ortonormal del espacio símplex $\mathcal{S}^D$ donde están definidos los datos originalmente, $\mathcal{B} = \{e_1, ..., e_{D-1}\}$. A partir de esa base se define $\mathbf{\Psi}$, una matriz $D-1 \times D$ donde la fila i es $clr(\mathbf{e}_i)$. Finalmente se calcula la transformación ILR

$$ilr(\mathbf{x}) = clr(x) \cdot \mathbf{\Psi}^\top \tag{2.4}$$

Es claro que dependiendo de la base que se haya utilizado obtendremos diferentes resultados.

Partición binaria secuencial Se abrevia SBP por sus siglas en inglés. En este último caso el conjunto de todas las componentes se divide en dos grupos, cada unos de estos dos grupos se subdividen en dos grupos, así sucesivamente hasta que finalmente solo quedan grupos de una sola componente. Las coordenadas cartesianas en esta nueva base se denominan balances. Esta partición de las componentes en subgrupos puede facilitar la interpretación ya que la agrupación de variables puede deberse algún tipo de relación de interés. Suele aplicarse esta transformación cuando se tiene algún tipo de conocimiento a priori que apunte a alguna subdivisión concreta y cuando el número de componentes no es demasiado alto, pues el número subparticiones se dispara, y con él el coste computacional, a medida que el número de componentes aumenta.

De manera más detallada lo que se haría sería dividir las D partes de la composición en dos subgrupos, $\mathbf{x_c}$ conteniendo c partes de la composición y $\mathbf{x_s}$ con las $D-c$ componentes restantes. Lo que buscamos es proyectar la composición general sobre el subespacio correspondiente a la subcomposición $\mathbf{x_c}$. Para ello se utiliza como matriz de proyección una matriz donde se codifica esa partición binaria secuencial. Veamos un ejemplo de como se construyen este tipo de matrices.

Supongamos una composición con D=5 partes y que hacemos una primera separación con $\mathbf{x_c} = (x_1, x_2)$ y $\mathbf{x_s} = (x_3, x_4, x_5)$.

Esta matriz signaria (signary matrix) codifica las particiones en subcomposiciones que se hacen en cada paso. La primera fila indica la división inicial que se hace, $\mathbf{x_c} = (x^1, x^2)$ por un lado y $\mathbf{x_s} = (x^3, x^4, x^5)$ por otro. En el segundo paso se divide el primer grupo en dos, como cada un de estos grupos queda conformado por una sola parte (x^1 y x^2 respectivamente) pasamos a las subdivisiones de $\mathbf{x_s}$ que quedan señaladas en los pasos 3 y 4 de manera análogo a como hicimos en los pasos anteriores. Para cada fila tenemos además las columnas r y s que simplemente indican el número de partes en cada subgrupo. Ignorando estas dos últimas columnas tendríamos una matriz signaria. A partir de ella se calcula la matriz de balances Ψ que necesitaríamos para computar las coordenadas de nuestros datos en esta nueva base. Para ello primeramente debemos darnos cuenta de que en 2.5 se divide en cada paso la composición total en 3 grupos, G_+, G_- y G_0, según si la parte tiene un 1, -1 o 0

Cuadro 2.5: Matriz signaria

Paso	x^1	x^2	x^3	x^4	x^5	r	s
1	1	1	-1	-1	-1	2	3
2	1	-1	0	0	0	1	1
3	0	0	1	1	-1	2	1
4	0	0	1	-1	0	1	1

Cuadro 2.6: Matriz Ψ correspondiente a matriz signaria 2.5

Paso	x^1	x^2	x^3	x^4	x^5
1	$\sqrt{\frac{3}{10}}$	$\sqrt{\frac{3}{10}}$	$-\sqrt{\frac{2}{15}}$	$-\sqrt{\frac{2}{15}}$	$-\sqrt{\frac{2}{15}}$
2	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0	0	0
3	0	0	$\frac{1}{\sqrt{6}}$	$\frac{1}{\sqrt{6}}$	$-\sqrt{\frac{2}{3}}$
4	0	0	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	0

respectivamente. Las partes de G_+ pasarán a ser $\frac{1}{r}\sqrt{\frac{rs}{r+s}}$, las de G_- pasarán a ser $\frac{-1}{s}\sqrt{\frac{rs}{r+s}}$ y las de G_0 serán 0. Siguiendo estas normas la matriz Ψ correspondiente a 2,5 sería

Al igual que ocurría en con la transformación ILR, cuando se habla de la transformación SBP, se está hablando de un conjunto de posibles transformaciones, no de una única (salvo en el caso de tener dos partes ya que entonces solo hay una posible partición binaria). Algunas diferencias con la transformación ILR es que no es necesario buscar una base ortonormal del símplex y que la idea tras esta transformación es bastante más intuitiva que en aquel caso. El inconveniente principal es que para un alto número de componentes este método no es viable.

2.3. Métodos de clasificación

Exploraremos tres vías distintas a la hora de plantear los criterios de clasificación. Primeramente, en una aproximación más afín al análisis multivariante clásico, estudiaremos el desempeño de varios de los criterios de clasificación supervisada más comunes como son el análisis discriminante lineal o la regresión logística entre otros. En esta primera etapa el objetivo no es hallar el mejor clasificador posible para cada algoritmo ya que, precisamente por su extendido uso, hay en la literatura innumerables posibilidades de ajustes y refinamientos que escapan a nuestros objetivos. Lo que buscamos en este trabajo es dar una muestra de como un enfoque composicional puede mejorar el rendimiento de todos estos algoritmos que como venimos diciendo son de uso tan común. Para mostrar esta mejoría que se produce al utilizar herramientas que habilitan los datos composicionales lo que se hará es comparar los modelos resultantes al utilizar los datos de los GPs en bruto frente a los modelos en los que previamente hemos aplicado alguna de las transformaciones descritas en la sección 2.2.4(concretamente la transformación ILR). Juzgamos útil presentar esta comparación directa entre métodos desde ambos enfoques porque, al cambiar solo los propios datos con los que se calcula cada uno de los criterios de

clasificación, conseguimos que las diferencias de rendimiento solo puedan ser explicadas por el hecho de haber utilizado una transformación composicional, dejando claro que es una herramienta útil y a tener en cuenta para analizar datos de estructura similar. Además, en muchas investigaciones, por diferentes motivos, no siempre se invierten la mayoría de recursos de una investigación en buscar el criterio de clasificación más sofisticado posible, basta con encontrar uno suficientemente bueno. Así pues, presentando mejorías en criterios de uso tan extendido y teniendo en cuenta su sencilla implementación, pretendemos dejar patente que, incluso desde un punto de vista muy pragmático, el uso de los datos composicionales debería ser una herramienta más a tener en consideración.

Posteriormente se tratará una aproximación diferente al problema, en vez de tratar de buscar una traducción de los métodos habituales a la situación que plantean los datos composicionales se desarrolla un algoritmo de clasificación que trabaja dentro del mundo composicional. Este algoritmo se conoce como *selbal* Rivera-Pinto et al., 2018 y se basa en la utilización de balances de las partes de la composición. Utilizando este algoritmo pretendemos dejar patente que el análisis composicional va más allá de buscar como adaptar métodos ya conocidos a esta nueva situación, que desde luego es algo útil y que proporciona herramientas versátiles y de más sencilla implementación, sino que también es viable estudiar métodos algo más novedosos.

2.3.1. Transformación ILR

En este punto, planteamos varios de los clasificadores más comunes a la hora de afrontar un problema de clasificación supervisada binaria. Tenemos dos capas en las que operar, por un lado tenemos el propio algoritmo clasificador y por otro qué variables son utilizadas para entrenar los distintos modelos. Los algoritmos que estudiaremos serán los siguientes (acompañados de las abreviaturas que se usarán en adelante para referirnos a cada uno de ellos:

- Análisis discriminante lineal (LDA)
- Análisis discriminante cuadrático (QDA)
- Regresión Logística (RL)
- Random Forest (RF)
- Naive Bayes (NB)

Y los distintos conjuntos de variables que serán usados en cada uno de estos modelos los podemos agrupar en dos subgrupos. El primer subgrupo englobará a los casos en los que se utilizan los datos de los cromatogramas en bruto, es decir, aquellos casos en los que usamos los GPs directamente. El segundo subgrupo consiste en los casos en los que previamente hemos transformado los datos de los GPs, concretamente utilizaremos las coordenadas proporcionadas por la transformación ILR. A su vez, tanto cuando usemos GPs como cuando usemos las coordenadas ILR planteamos las siguientes situaciones y sus correspondientes abreviaturas que se usarán en el resto del texto entre paréntesis:

- Totalidad del cromatograma. Utilizando todos los GPs o todas las coordenadas ILR (*NT-GP* y *ILR-GP* respectivamente)
- Totalidad del cromatograma más variables demográficas. Añadiendo las variables de sexo, edad e índice de masa corporal de cada individuo al caso anterior. (*NT-GPdemo* y *ILR-GPdemo* respectivamente)
- Parte del cromatograma. Seleccionando solo algunos de los GPs o algunas de las coordenadas ILR. (*NT-GPV* y *ILR-GPV* respectivamente)
- Parte del cromatrograma más sociales. Añadiendo las variables demográficas ya mencionadas al caso anterior. (*NT-GPVdemo* y *ILR-GPVdemo* respectivamente)

Selección de variables

. Antes de explicar los distintos algoritmos de clasificación, explicaremos por qué se han establecido los subgrupos de variables que se han descrito anteriormente. Los grupos *GP* y *ILR* se establecieron porque son los que sirven para comparar de manera más directa las bondades de aplicar las transformaciones composicionales, se eligió la transformación ILR porque es la que suele ser más versátil computacionalmente ya que elimina completamente la restricción de suma constante, algo que no se consigue con la transformación *CLR*, lo cuál supone un problema para algunos modelos como por ejemplo LDA. Además no requiere de una subdivisión constante de las partes como sí requiere la transformación SBP 2.2.4, lo cuál supondría un coste computacional y una complejidad añadida que no valoramos que fuera compensada por una notable mejora en el rendimiento del clasificador. La transformación SBP se suele utilizar cuando ya se tiene a priori algún tipo clasificación de las partes, agrupación de la que no disponíamos en nuestro caso. Se podría llegar a plantear el subdividir los GPs en categorías según su complejidad, pero no llegamos a encontrar ningún criterio especialmente claro para ello. Además se utilizan todos los GPs sabiendo que es la situación donde la colinealidad de las variables es total, es decir donde más debería notarse la ventaja de aplicar alguna de las transformaciones, y como se utilizan todos los GPs, lo más equiparable sería utilizar todas las coordenadas una vez aplicada una transformación. En cuanto a las variables demográficas, se han utilizando edad, sexo e índice de masa corporal porque de todas las variables de las que disponíamos se acordó con el Dr. Francisco Gude que serían las más relevantes. También se optó por ellas ya que otro objetivo de este trabajo es hacer notar la sencillez y aplicabilidad de estas transformaciones, así que juzgamos interesante dar una muestra de como se puede aplicar a variables composicionales sin dejar de lado variables tan habituales como estas.

En cuanto a los casos donde solo usamos una parte de los GPs (o de las coordenadas ILR) el método de selección que se utilizó fue el implementado en el paquete VSURF (Genuer et al., 2022). Para un descripción más detallada del procedimiento de este algoritmo se puede consultar su manual disponible en CRAN. Este paquete incluye varias funciones con distintos propósitos, nosotros utilizamos una de sus funciones principales, *VSURF*. Esta función, basada en bosques aleatorios (RF por sus siglas en inglés) consta de tres pasos. El primer paso ("thresholding step") trata de eliminar las variables irrelevantes. En el segundo paso (ïnterpretation step"), trata de seleccionar las variables relacionadas con la variable respuesta con fines de interpretación. En el último paso trata de refinar esta selección de variables cuando el objetivo es la predicción. Esta función devuelve por tanto dos conjuntos, el del paso 2 para interpretar y el del paso 3 para predecir. En la práctica no notamos diferencias destacables a la hora de usar uno u otro.

Cabe destacar que cuando se usaba este procedimiento automático de selección de variables, la selección era mucho más permisiva cuando se aplicaba sobre los GPs que cuando se hacía sobre las coordenadas ILR. Llegando a quedarse con 40 de las 46 GPs en el segundo paso pero con tan solo 3 cuando se aplicaba sobre las coordenadas ILR. Aún sin haber encontrado una explicación definitiva creemos que es en parte debida a la colinealidad de las variables, y aunque muestra una vez más lo beneficioso de aplicar estas transformaciones, más estudio sería necesario para saber exactamente el motivo de esta disparidad. Finalmente se optó por utilizar el grupo de variables resultante del tercer paso porque nuestro propósito con la clasificación es predecir y porque en este caso sí que se obtenía una selección reducida incluso al tratar con los GPs directamente.

En función de si se ha aplicado o no la transformación ILR, de si se han utilizado o no las variables demográficas y de la utilización o no del paquete la función *VSURF* se establecen los siguientes grupos de variables junto a sus correspondientes abreviaturas.

- Todos los GPs, sin aplicar transformaciones (NT-GP)
- Todos los GPs y las variables demográficas edad, sexo, índice de masa corporal (NT-GPdemo)
- Los GPs seleccionados por VSURF (NT-GPV)

- Los GPs seleccionados por VSURF más las variables demográficas (NT-GPVdemo)
- Todas las coordenadas ILR (ILR-GP)
- Todas las coordenadas ILR y variables demográficas (ILR-GPdemo)
- Coordenadas ILR seleccionadas por VSURF (ILR-GPV)
- Coordenadas ILR seleccionadas por VSURF y variables demográficas (ILR-GPVdemo)

Ahora veamos en detalle de cada uno de los algoritmos de clasificación empleados.

Análisis discriminante

Se basa en la búsqueda de una combinación lineal de las variables predictoras que separe de la mejor manera posible las dos clases de la variable respuesta. Podemos expresar dicha combinación lineal como:

$$Z = w_1x_1 + w_2x_2 + ... + w_px_p \tag{2.5}$$

Este criterio de separación se formaliza mediante la siguiente expresión:

$$S(\mathbf{w}) = \frac{\sigma^2_{inter}}{\sigma^2_{intra}} = \frac{(\mathbf{w}\cdot\mu_\mathbf{1} - \mathbf{w}\cdot\mu_\mathbf{0})^2}{\mathbf{w}^T\Sigma_1\mathbf{w} + \mathbf{w}^T\Sigma_0\mathbf{w}} = \frac{(\mathbf{w}\cdot(\mu_\mathbf{1} - \mu_\mathbf{0}))^2}{\mathbf{w}^T(\Sigma_1 + \Sigma_0)\mathbf{w}} \tag{2.6}$$

Esta expresión se maximiza cuando

$$\mathbf{w} \propto (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0) \tag{2.7}$$

El análisis discriminante lineal trabaja bajo varias asunciones:

1. La distribución de $p(\mathbf{x}|y = 0)$ y $p(\mathbf{x}|y = 1)$ siguen una distribución normal, con parámetros (μ_0, Σ_0) y (μ_1, Σ_1).
2. Independencia lineal de las variables explicativas.
3. Individuos muestreados de manera aleatoria e independiente.
4. Homocedasticidad en las clases de la variable respuesta.

Si bien el cumplimiento de la asunción de normalidad hace que los resultados de LDA sean más estables, se ha comprobado como no es una condición absolutamente necesaria para obtener resultados suficientemente buenos (Li et al., 2006).

La última asunción implica que $\Sigma_0 = \Sigma_1 = \Sigma$ permite simplificar 2.7 a

$$\mathbf{w} \propto (\Sigma)^{-1}(\mu_1 - \mu_0) \tag{2.8}$$

Cuando no se hace esta asunción de homocedasticidad se considera análisis discriminante cuadrático (QDA). Este algoritmo presenta un pequeño inconveniente y es que el hecho de que tengamos más variables que individuos en algún grupo imposibilita el cálculo de las matrices Σ_0 y Σ_1, por ellos, los modelos calculados basados en QDA que trataremos son solo los que hayan utilizado como variables NT-GPV, NT-GPVdemo, ILR-GPV e ILR-GPVdemo ya que la previa selección automática de variables del algoritmo reduce suficientemente el número de covariables como para permitir el cálculo de las matrices.

Regresión Logística Múltiple

Tenemos una variable respuesta dicotómica que codifica con 1 a los pacientes que han fallecido por cáncer y con 0 los que no. La relación entre las variables predictoras, que pueden ser tanto continuas como discretas, y la variable respuesta se establece a través de una función link. Existen distintas opciones para tomar como función link, una de las más comunes y por la que hemos optado es la función logit, que tomaría la siguiente forma.

$$p(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\beta \cdot \mathbf{x}}}{1 + e^{\beta \cdot \mathbf{x}}} \tag{2.9}$$

Donde $\beta = (\beta_0, \beta_1, ..., \beta_k)$ es el vector de coeficientes del modelo, $\mathbf{X} = (1, X_1, X_2, ..., X_k)$ el conjunto de variables independientes precedidas de un 1 y $\beta \cdot \mathbf{X}$ el producto escalar entre ambos vectores. Dado un individuo concreto, el resultado de evaluar 2.9 según el valor que tomen las variables independientes en ese individuo determinará en que grupo se le clasifica. La clasificación surge de evaluar si la puntuación que el individuo obtiene según el modelo supera o no un determinado umbral. Dicho umbral se suele establecer por defecto en 0.5 aunque se puede cambiar para ajustar el número de falsos positivos y falsos negativos. El modelo logístico es ampliamente utilizado por su sencillez, flexibilidad ya que nos permite adaptarnos a distintas tesituras cambiando la función link, y por su interpretabilidad, ya que en casos sencillos los coeficientes pueden dar una clara visión de la relación entre las variables independientes y la variable respuesta. Sin embargo, el objetivo de este trabajo no es encontrar el mejor método de clasificación posible sino dejar de manifiesto como el simple hecho de utilizar los datos composicionales hace que partamos de una mejor posición para llevar a cabo cualquier análisis estadístico. Por supuesto, los modelos que veamos que utilizan los datos composicionales son susceptibles de un mayor refinamiento a posteriori para dar con los hiperparámetros del modelo que conducen al modelo óptimo.

Bosque aleatorio

Antes de explicar brevemente en qué consiste este algoritmo es conveniente entender qué son los árboles de decisión, pues son los ladrillos sobre los que se cimienta un bosque aleatorio (RF).

Un árbol de decisión se compone de dos partes, nodos y hojas. Cada nodo representa un criterio de partición de los datos, que una variable continua supere determinado umbral o que una variable categórica sea de una categoría concreta, por ejemplo. Cada nodo genera dos hojas y asigna cada individuo a una u otra hoja en función del criterio del nodo. A su vez, de cada hoja puede surgir un nuevo nodo volviendo así a la situación inicial. El objetivo de cada nodo es separar las clases de la variable respuesta lo mejor posible la submuestra que se encuentra en la hoja en la que está asociado (en el caso del primer nodo sería la muestra total).

Cuando una hoja ya solo tiene individuos de una clase no hace falta seguir sacando nodos de ella, sin embargo no siempre se llega a una clasificación perfecta para parar el desarrollo del árbol, ya que con datos reales esa meta solo se cumpliría al llegar a árboles excesivamente complejos. Criterios de parada hay varios, puede establecerse desde un principio un número máximo de nodos o se puede determinar algún criterio que establezca cuando el árbol ya hace una clasificación satisfactoria y por tanto no sea necesario continuar su crecimiento mediante más nodos.

Una vez establecido el concepto de árbol de decisión, la idea del algoritmo RF es muy intuitiva. Como su nombre indica consiste en utilizar multitud de árboles de decisión de manera simultánea. Para ello, primeramente se desarrollan varios árboles que pueden diferir tanto en número de nodos como en los criterios utilizados en cada nodo. Una vez desarrollados los árboles, RF clasifica cada individuo atendiendo a dónde lo clasifican la mayoría de árboles. La idea es conseguir compensar la falta de precisión de cada uno de los árboles a nivel individual mediante el consenso de la mayoría así,

individuos que por sus características serían clasificados con menor acierto en algunos árboles en específico, son clasificados de manera correcta por el RF ya que son bien clasificados en el resto de árboles.

Los RF son una herramienta muy potente, versátil y profunda pero, de nuevo, la idea de este trabajo no es entrar en profundidad en cada uno de los métodos y explorar su máximo potencial, sino ver como el uso de los datos composicionales puede reportar resultados positivos en variedad de escenarios distintos.

Naive Bayes

La regla de clasificación que utiliza este algoritmo se basa en el teorema de Bayes. Dicho teorema tiene una formulación más genérica pero, adaptándolo a nuestra situación, donde tenemos una variable respuesta categórica Y junto a una serie de variables independientes $\mathbf{X} = (X_1, ..., X_k)$ que queremos usar como predictoras, se puede formular de la siguiente manera:

$$p(Y = C_i|\mathbf{X} = \mathbf{x}) = \frac{p(C_i)p(\mathbf{X} = \mathbf{x}|Y = C_i)}{p(\mathbf{X} = \mathbf{x})} \tag{2.10}$$

La razón de utilizar esta aproximación es que estas probabilidades no son conocidas en un estudio real, por tanto han de aproximarse con los datos de los que dispongamos. Cuando el número de variables independientes es alto o cuando alguna de estas variables puede tomar gran cantidad de valores (como ocurre cuando una variable es real por ejemplo), aproximar $p(Y = C_i|\mathbf{X} = \mathbf{x})$ a partir de los datos es complicado debido a la cantidad posibles opciones para el vector $\mathbf{x}$.

Utilizando la regla de la cadena de la probabilidad condicionada sucesivamente podemos reescribir el numerador de 2.10 de la siguiente forma

$$p(C_i)p(\mathbf{X} = \mathbf{x}|Y = C_i) = p(C_i, \mathbf{x}) = p(x_1|x_2, ..., x_k, C_i)p(x2, ..., x_k, C_i) = p(x_1|x_2, ..., x_k, C_i)p(x2|x_3..., x_k, C_i)p(x_3, ..., x_k, C \tag{2.11}$$

El siguiente paso, el que le da la parte de *Naive* al nombre del clasificador, es asumir que las variables predictoras son independientes entre sí, de manera que muchas de las probabilidades condicionadas utilizadas en 2.11 se simplifican rápidamente de la siguiente manera

$$p(x_j|x_{j+1}, ..., x_k, C_i) = p(x_j|C_i) \tag{2.12}$$

Esto hace que podamos expresar 2.11 como

$$p(C_i) \prod_{j=1}^{k} p(x_j|C_i) \tag{2.13}$$

A la hora de predecir la clase de un individuo conoceremos los valores de las variable predictoras y, en consecuencia, el denominador de 2.10 será realmente una constante ya que en ningún momento depende de la variable respuesta. Esto hace que podamos expresar 2.10 como

$$p(Y = C_i|\mathbf{X} = \mathbf{x}) \propto p(C_i) \prod_{j=1}^{k} p(x_j|C_i) \tag{2.14}$$

Finalmente, el clasificador que se propone es el siguiente

$$\hat{Y} = argmax_{i \in (1,...,M)} p(C_i) \prod_{j=1}^{k} p(x_j|C_i) \tag{2.15}$$

Consiste simplemente en asignar a cada individuo a la clase que maximice la expresión 2.14. A la hora de poner en práctica este clasificador se deben estimar distintos valores. Por un lado la probabilidad de cada clase, $p(C_i)$, se suele estimar como $\frac{\#individuosenlaclase C_i}{\#totalindividuosenlamuestra}$ aunque también se podría

establecer en otros valores propuestos en la literatura. Para estimar $p(X_j|Y = c_i)$ en los casos en los que sea X_j una variable continua, se suele optar por asumir normalidad en la distribución y estimar la media y varianza a partir de los individuos que pertenezcan a la clase C_i. Otra opción habitual es discretizar la variable en varios niveles y estimar en función de la cantidad de individuos que haya en cada nivel con la categoría de la variable respuesta previamente fijada. Estas no son las únicas opciones ya que se podría llegar a ser mucho más sofisticado e incluso llevar a cabo estimaciones no paramétricas de la distribución de estas probabilidades condicionadas, explorar con más detalle este amplio abanico de opciones se escapa de los objetivos de este trabajo. Nosotros nos decantamos por la opción de asumir normalidad.

Adicionalmente, para tratar el problema del desbalance de las clases, se en los algortimos LDA, QDA y NB se estableció la misma probabilidad a priori en sendas clases. Para los los modelos RF se realizo un sobreremuestreo del grupo de fallecidos por cáncer mediante la función *ovunsample* del paquete ROSE (Lunardon et al., 2014) que trataba de crear una muestra artificial donde ambas clases estuvieran balanceadas. Para los modelos de RL se asigno un peso a cada individuo en función de la clase a la que perteneciese, concretamente el peso era el inverso del tamaño del grupo al que pertenecía, de esta manera los individuos víctimas mortales del cáncer tienen un mayor peso que los otros.

2.4. Algoritmo *selbal*

El algoritmo *selbal* Rivera-Pinto et al., 2018 fue propuesto en el contexto de análisis de microbiomas. En su artículo, su objetivo es identificar firmas microbianas que puedan ser usadas en el diagnóstico, pronóstico o predicción de la respuesta terapéutica de individuos basándose en sus datos microbianos particulares.

Para realizar esta tarea de discriminación buscan un balance entre dos subgrupos del total de taxones que se tuvieron en cuenta a la hora de recoger actividad microbiana de los pacientes. El análisis de esos datos concretos requiere de un enfoque composicional Gloor et al., 2017 y este algoritmo fue diseñado con ello en mente, por ello creemos que es adecuado utilizarlo en nuestro caso particular. Además, una de las limitaciones que presenta el algoritmo es la necesidad de eliminar todos los ceros de de las muestras. Esta limitación no es exclusiva de este algoritmo, sino que es una problemática común a muchos de los métodos típicos del análisis composicional. Esto se debe a que, como vimos anteriormente, es común trabajar en algún punto con logaritmos o cocientes entre partes, funciones que no están definidas en 0. En el caso del algoritmo *selbal*, el uso de los balances, es lo que impide trabajar ante la presencia de ceros. El tratamiento de ceros en datos composicionales es un tema sobre el que aún se realiza investigación a cerca de los efectos que tienen distintas técnicas para atajar el problema que conllevan. Existen enfoques más sencillos Greenacre, 2021 y otros algo más sofisticados Templ, 2021. Pese a lo interesante de la problemática, debido a que nuestros datos no presentan ceros, no entraremos en más profundidad. Cabe decir que *selbal* incluye medidas para tratar la presencia de ceros, pero no hay mejor medida para remediar un problema que no tener el problema desde un inicio.

El funcionamiento del algoritmo *selbal* se puede resumir como un algoritmo avaricioso en el que se parte de un balance con solo dos partes y se van añadiendo partes al balance hasta que se cumple un determinado criterio de parada. En cierta medida se parece a un algoritmo discriminante lineal avaricioso, solo que aquí, en lugar de añadir variables a una ecuación lineal, se añaden a un balance. Veremos más adelante que este proceso de añadir partes al balance sigue una estructura muy similar a lo que se hace en el discriminante lineal. Pero antes de ver el funcionamiento de este algoritmo debemos definir exactamente que es un balance.

Si tenemos una composición $\mathbf{X} = (X^1, X^2, ..., X^D)$, dos subgrupos de índices disjuntos, I_+, I_- cada uno de tamaño k_+ y k_- respectivamente y que definen dos subcomposiciones, $\mathbf{X}_+ = \{X^i \ i \in I_+\}$ y

$\mathbf{X}_- = \{X^i \; i \in I_-\}$. El balance entre estas dos subcomposiciones se define como:

$$\mathbf{B}(\mathbf{X}_+, \mathbf{X}_-) = \sqrt{\frac{k_+ k_-}{k_+ + k_-}} log \frac{(\prod_{i \in I_+} X^i)^{1/k_+}}{(\prod_{j \in I_-} X^j)^{1/k_-}} \tag{2.16}$$

En la expresión 2.16 se puede desarrollar el logaritmo para obtener una más amable desde el punto de vista computacional.

$$B(\mathbf{X}_+, \mathbf{X}_-) \propto \frac{1}{k_+} \sum_{i \in I_+} logX_i - \frac{1}{k_j} \sum_{j \in I_j} logX_j \tag{2.17}$$

A partir de la expresión 2.17 se puede intuir mejor el parecido entre un modelo lineal tradicional y la aproximación composicional de trabajar con balances. Las diferencias con un modelo lineal son, por un lado, que aquí estamos trabajando con las variables transformadas por el logaritmo (recordemos lo oportuno que resulta no tener ceros en nuestros datos) y que los coeficientes de la ecuación presentan la restricción adicional de que suman 0. Una vez establecido el concepto de balance de manera rigurosa, veamos los pasos que se toman en el algoritmo.

Paso 1:
Se analizan todos los balances posibles utilizando solo dos componentes, es decir todos los balances de la forma:

$$B_{ij} = \sqrt{\frac{1}{2}}[log(X^i) - log(X^j)] \; con \; i, j \in \{1, ..., D\} \; i \neq j$$

Denotando por Y la variable respuesta dicotómica que queremos modelar, *selbal* ajusta el siguiente modelo de regresión para cada balance

$$logit(Y) = \beta_0 + \beta_1 B_{ij} + \gamma' Z \tag{2.18}$$

donde $\mathbf{Z}$ representa una serie de covariables no composicionales (en nuestro caso podría ser el IMC, sexo y edad). El balance que maximice el criterio de optimización será denotado por B_1. Al tratarse de un problema de clasificación dicotómica el criterio de optimización empleado es ver que modelo maximiza el valor AUC Paso n:
Para n¿1 y hasta que se verifique el criterio de parada y dado el balance definido en el paso anterior:

$$B^{(s-1)} \propto \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} log(X_i) - \frac{1}{k_j^{(s-1)}} \sum_{j \in I_j^{(s-1)}} log(X_j) \tag{2.19}$$

para cada parte $X^p \quad p \notin [I_+^{(s-1)} \cup I_-^{(s-1)}]$ se calculan dos nuevos balances. El primero resulta de añadir $log(X^p)$ a la parte positiva de 2.19

$$B_p^{(s^+)} \propto \frac{1}{k_+^{(s-1)+1}}[\sum_{i \in I_+^{(s-1)}} log(X^i) + log(X^p)] - \frac{1}{k_j} \sum_{j \in I_j^{(s-1)}} log(X^j)$$

y el segundo surge de añadirlo a la parte negativa

$$B_p^{(s^-)} \propto \frac{1}{k_+^{(s-1)}} \sum_{i \in I_+^{(s-1)}} log(X^i) - \frac{1}{k_j^{(s-1)} + 1}[\sum_{j \in I_j^{(s-1)}} logX^j + log(X^p)]$$

Y una vez calculados estos dos balances para cada una de las partes restantes, X^p, se estima un modelo como el de 2.18

$$logit(Y) = \beta_0 + \beta_1 B_{ij} + \gamma' Z$$

Aquel balance que maximice el criterio de optimización será seleccionado como balance $B^(s)$
Criterio de parada:
Hay dos criterios de parada. Cuando la mejora obtenida en el criterio de parada al utilizar un nuevo balance no supera cierto umbral o cuando sea alcanzado un número máximo de componentes especificado previamente.

Este algoritmo no garantiza encontrar el máximo global ya que podría darse el caso en el que el máximo global se de utilizando el balance $\frac{X^2X^3}{X^4X^5}$ pero que de los balances de dos partes iniciales el mejor AUC se obtenga con el balance $\frac{X^1}{X^2}$ y por tanto nunca llegaremos al verdadero balance óptimo porque siempre estaremos incluyendo la parte X^1 cuando no es necesaria. Para paliar en cierta medida esta incertidumbre se recurre a la validación cruzada, de esta manera, aunque seguimos sin garantizar haber obtenido el balance óptimo global, podemos comprobar la robustez del resultado y tener algo más de confianza a la hora de tratarlo como si fuera el punto óptimo.

2.5. Comparación

En 2.3.1 se proponen, cruzando los distintos algoritmos de clasificación con los distintos grupos de variables usadas para entrenarlos modelos, un total de 36 modelos. Para evaluar y comparar el rendimiento de cada uno de ellos se utilizaron distintos índices y medidas.

En problemas de clasificación supervisada una de las herramientas más básicas empleadas a la hora de evaluar el desempeño de los clasificadores es la tablas de confusión. Siguiendo la siguiente notación:

- VP Número de verdaderos positivos, enfermos clasificados como tales
- VN Número de verdaderos negativos, sanos clasificados como tales
- FP Número de falsos positivos, sanos clasificados como enfermos
- FN Número de falsos negativos, enfermos clasificados como sanos

		Criterio Real	
		Enfermos	No enfermos
Diagnóstico	Positivos	VP	FP
	Negativos	FN	VN

Una tabla de confusión típica en un contexto sanitario como en el que estamos podría ser la siguiente:

A partir de esta tabla se pueden calcular una serie de índices que tratan de resumir la calidad de los modelos de clasificación. Se explican brevemente a continuación cada una de ellas.

Exactitud

La exactitud se calcula como $\frac{VP+VN}{VP+VN+VP+VN}$ y mide la proporción de aciertos totales. Aunque da un visión general del clasificador, ante clases desbalanceadas (como es nuestro caso) pueden dar una visión engañosa.

Precisión
La precisión se calcula como $\frac{VP}{VP+FP}$ y mide el acierto de las predicciones positivas, es útil cuando los falsos positivos son altamente costosos.

Sensibilidad
La sensibilidad se calcula como $\frac{VP}{VP+FN}$, mide la proporción de positivos reales que fueron correctamente identificados. En las situaciones en las que se prefiera un falso positivo a un falso negativo (como es habitual en el contexto médico) el incremento de este índice es objetivo principal a la hora de establecer mejores modelos.

Especificidad
La especificidad se calcula como $\frac{VN}{VN+FP}$, calcula la proporción de verdaderos negativos que son correctamente identificados. Aunque es cierto que en el contexto médico la sensibilidad suele ser foco principal de atención, la especificidad no debe ser descuidada pues a la hora de aplicar tratamientos agresivos (como la quimioterapia) se debe tener la máxima certeza posible de no estar tratando un falso positivo.

F1-Score
Se calcula como $2 \cdot \frac{Precisión \cdot Sensibilidad}{Precisión+Sensibilidad}$. Se suele emplear en situaciones de clases desbalanceadas. Combinando la sensibilidad y la precisión penaliza de igual manera los falsos positivos que los falsos negativos.

A partir de estos conceptos se puede ir un paso más allá mediante la utilización de lo que se conoce como curvas ROC (Receiver Operating Characteristic) y sus respectivos AUC (Area Under the Curve). Las curvas ROC son una herramienta muy común a la hora de evaluar clasificadores binarios supervisados (Hoo et al., 2017). Para entender como se calculan las curvas ROC, debemos entender como funcionan estos clasificadores. Es claro cual es el objetivo, dado un individuo, el clasificador debe asignarlo a alguna de las dos categorías preestablecidas de la variable respuesta atendiendo a los valores que presente el individuo en las distintas variables predictoras usadas para entrenar el modelo.

Estos modelos sin embargo, no dan directamente la clasificación del individuo, sino que lo que hacen es estimar las probabilidades de que pertenezca a cada una de las clases. Una vez establecido un umbral de probabilidad para una de las clases, p_1, un individuo es asignado a dicha clase en caso de que el individuo supere el umbral, si no se supera este umbral se asigna a la otra. Al haber solo dos clases, una vez establecido p_1 queda implícitamente determinado también el umbral de la segunda clase como $p_2 = 1 - p_1$. Hay casos en los que no se obliga a que los umbrales de cada clase sumen 1, estos casos permiten que el clasificador deje sin catalogar individuos en los que el criterio de discriminación no sea muy claro, nosotros no contemplamos esa opción. De esta manera queda mostrado que la regla de clasificación no depende solo de los datos o del algoritmo utilizado, sino que, en última instancia, la calidad del clasificador se ve fuertemente influenciada por el valor que se establezca para el umbral. Da igual la regla de clasificación usada, si estableciésemos $p_1 = 1$ como umbral para la clase donde la variable respuesta es positiva, sería prácticamente imposible obtener un falso positivo, pues no clasificaríamos prácticamente ningún individuo como positivo. Una vez establecido que este umbral influye de manera determinante en la calidad de los clasificadores, la idea tras el cálculo de una curva ROC es muy sencilla. Consiste en establecer diferentes umbrales, desde los más permisivos a los más exigentes, de manera que para cada uno de estos umbrales se calculan la sensibilidad y especificidad del modelo y se representa en una curva donde la coordenada X es 1-especificidad y la coordenada Y la sensibilidad. Nosotros hemos optado por representar en el eje X la especificidad directamente y por ello el el eje horizontal se recorre de izquierda a derecha de manera descendente. Además, tanto especificidad como sensibilidad las expresamos en porcentajes.

A la hora de interpretar una curva ROC de manera visual tenemos dos claves. La primera es que cuando más se aproxime a la esquina superior izquierda del gráfico mejor será el clasificador, ya que el

vértice de esa esquina representaría una especificidad y unas sensibilidad del 100 % es decir, sería un clasificador perfecto. La segunda diagonal que une el punto 100 % de especificidad y 0 % de sensibilidad con el punto 0 % de especificidad y 100 % de sensibilidad representa a un clasificador aleatorio, aquel que asigna a cada individuo con igual probabilidad a una u otra clase sin tener en cuenta ninguna variable predictora, cuanto más se aleja de esta línea (por encima) una curva ROC mejor será el clasificador que representa y al revés, si en algún momento pasa por debajo es que se trata de un clasificador pésimo. Para cuantificar como rinde un clasificador de manera global se suele utilizar el valor AUC, dicho valor representa el área que se encuentra por debajo de la curva. El AUC del clasificador aleatorio sería de 0.5 y por tanto cualquier clasificador con un valor por de debajo 0.5 se interpreta como que no tiene valor discriminatorio. Para los casos en los que se encuentra por encima no hay una escala fija aunque es habitual establecer en 0.75 el mínimo para que un clasificador se considere que pueda tener cierto valor clínico (Fan et al., 2006). En cualquier caso cuanto más próximo a 1 mejor.

A la hora de calcular todos los valores necesarios para obtener las curvas ROC necesitamos entrenar los modelos y evaluarlos. La práctica habitual es dividir los dos datos en dos submuestras, una primera submuestra que servirá de entrenamiento y una segunda submuestra que servirá de testeo. Esto se hace porque si utilizásemos los mismos datos que hemos utilizado para estimar el modelo para testearlo correríamos el riesgo de obtener unos resultados sesgados hacia la muestra y que realmente los modelos no fuesen tan buenos como aparentan de cara a tratar datos completamente nuevos, que es al final el objetivo de estos clasificadores. Una práctica habitual es establecer un porcentaje arbitrario, entorno al 20-30 % (Gholamy et al., 2018) y reservar dicho porcentaje de la muestra para testeo, utilizando el restante para el entrenamiento del modelo. Sin embargo nosotros hemos optado por utilizar otro método para validar los modelos. Hemos recurrido al método Leave One Out Cross Validation (LOOCV). La idea de esta manera de proceder es recorrer todos y cada uno de los individuos de la muestra, para cada individuo se toma el total de la muestra menos al individuo y se establece este subconjunto del total como set de entrenamiento. Una vez entrenado el modelo con este set se evalúa sobre el propio individuo, el conjunto de entrenamiento estaría formado única y exclusivamente por el individuo que se extrajo de la muestra original.

De manera algo más detallada el procedimiento será el siguiente.

1. Dada la muestra $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$ se establece como conjunto de entrenamiento $train = (\mathbf{x_1}, ..., \mathbf{x_{i-1}}, \mathbf{x_{i+1}}, ..., \mathbf{x_n})$ y como set de testeo $test = (\mathbf{x_i})$
2. sobre $train$ se entrena el modelo y se calcula $\hat{y}_i$, una predicción para la variable respuesta del individuo i. Concretamente guardaremos la probabilidad que le asigna el modelo a pertenecer a la categoría de positivo y no la clasificación en sí.
3. Repetimos los dos pasos anteriores hasta haber recorrido todos los individuos una sola vez. Denotamos por $\hat{\mathbf{Y}}$ el vector que guarda todas las probabilidades asignadas a cada individuo.
4. A partir del vector $\hat{\mathbf{Y}}$ y del vector $\mathbf{Y}$, el vector que contiene los valores reales de las variables respuestas de todos los individuos calculamos las curvas ROC.
5. De cada curva ROC obtenemos el índice Youden y lo utilizamos para establecer a que clase se asigna cada individuo. Una vez acabada la asignación para todos los individuos se compara con el vector $\mathbf{Y}$ y se calcula la tasa de acierto

Como hemos establecido anteriormente, dependiendo del umbral que se establezca se obtienen distintos criterio de clasificación, pero para calcular los índices que describíamos al principio necesitamos establecer un umbral para fijar el criterio de clasificación, determinar los falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos para, a partir de ellos, computar los diferentes índices. Para establecer dicho umbral se empleó el índice de Youden. Este índice cálcula el punto de curva ROC que maximiza la expresión $J = sensibilidad + especificidad - 1$. Y como umbral se establece aquel correspondiente a ese punto de la curva ROC.

Capítulo 3

Resultados

3.1. Clasificadores clásicos

Empezamos mostrando las curvas ROC de los modelos clásicos de clasificación supervisada para mostrar de una manera visual el desempeño general de cada uno de estos modelos. Posteriormente se muestran las tablas con los distintos índices de evaluación de los modelos explicados en el apartado 2.3.1. En este apartado nos resulta de especial interés dejar constancia de la mejora que supone el hecho de emplear la transformación ILR, con el objetivo de mostrar los potenciales beneficios que supone este procesado de los datos previo al cálculo de los modelos a cambio de un coste muy reducido pues la transformación ILR se computa rápidamente a través de la función $ilr()$ que vimos en 2.2.4. Se puede apreciar en la figura 3.1 como a través de casi todos los modelos, el hecho de aplicar la transformación ILR repercute en un mayor desempeño global del clasificador, siendo la diferencia entre NT-GPV e ILR-GPV con el modelo QDA, figura 3.1b, especialmente notable. En la figura 3.2 se presentan las curvas ROC asociadas a los modelos de regresión logística. De nuevo los modelos donde se ha aplicado la transformación ILR tienen un mejor rendimiento que los que se han calculado sobre los GPs directamente a excepción de la comparación del modelo NT-GPV e ILR-GPV (en azul). En este caso, si bien sus respectivas curvas ROC son parejas, si que la curva de NT-GPV parece dominar a la de ILR-GPV la mayor parte del tiempo. Como en QDA, el caso en el que se han seleccionado variables de manera automática y se han añadido las variable demográficas (naranja) parece ser el que muestra una mayor mejoría al pasar a utilizar la transformación ILR, aunque en el caso en el que no se han seleccionado variables ni añadido las demográficas (negro) también presenta una clara mejoría. En el caso de los modelos basados en bosques aleatorios, figura 3.3, cabe destocar por un lado el bajo desempeño general de los modelos. Es especialmente sorprendente ya que el paquete VSURF precisamente se base en bosques aleatorios para hacer la selección de variables así que cabría esperar que, al menos si nos centrásemos solo en los casos en los que se ha aplicado esta selección automática de variables, se diese lugar a un mejor clasificador respecto al resto de algoritmos. Finalmente, en la figura 3.4 tenemos las curvas rocROC de todos los modelos calculados a partir del método de Naive-Bayes. De nuevo, atendiendo a estas curvas, el modelo correspondiente a ILR-GPVdemo (naranja continuo) parece ser el que mejor clasificador global. También es notable que en los casos donde más se nota una mejoría al pasar a utilizar la transformación ILR son el caso de NT-GP frente a ILR-GP (negro) y NT-GPVdemo comparado con ILR-GPVdemo.

Otro aspectos interesantes surgen al comparar los modelos siguiendo dos nuevos criterios de emparejamiento. El primero sería realizar la comparativa entre modelos que solo se diferencian a la hora de emplear o no las variables demográficas, de manera visual, sería comparar en las gráficas las curvas negras frente a las rojas o las azules con las naranjas). El segundo criterio sería comparar los modelos que solo difieran en el uso o no de la selección automática provista por VSURF. Atendiendo al primer criterio, sería esperable que los modelos que incluyesen las variables demográficas fueran, al menos en

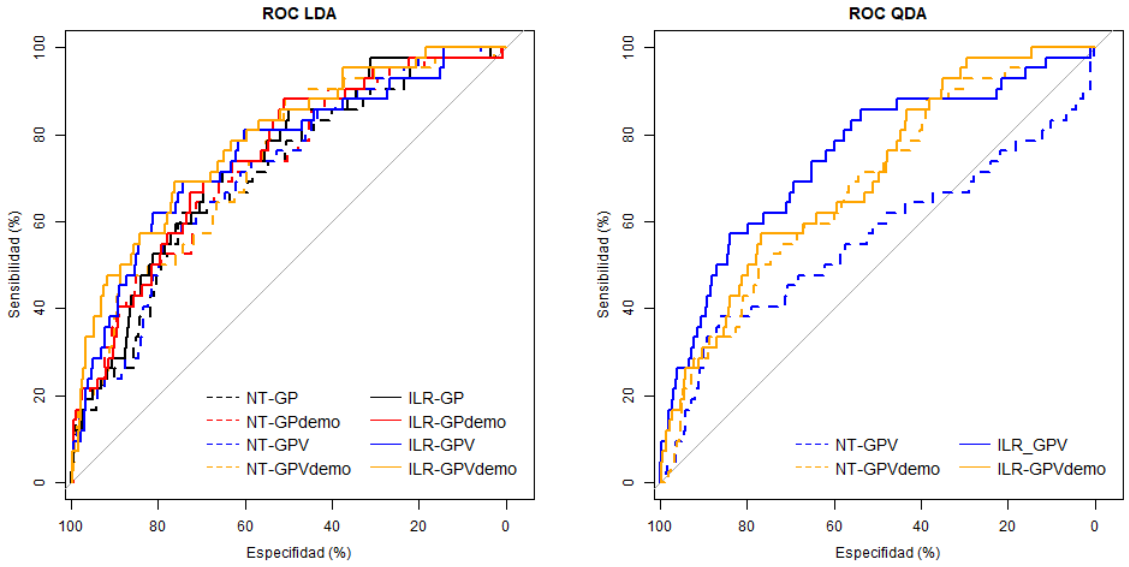
	Modelos LDA							
Criterio	NT-GP	ILR-GP	NT-GPdemo	ILR-GPdemo	NT-GPV	ILR-GPV	NT-GPVdemo	ILR-GPVsocio
Exactitud	0.7455	0.6953	0.7112	0.7250	0.7726	0.7422	0.5188	0.7614
Precisión	0.9857	0.9875	0.9859	0.9871	0.9845	0.9883	0.9934	0.9886
Sensibilidad	0.7492	0.6954	0.7131	0.7267	0.7784	0.7437	0.5085	0.7634
Especifidad	0.6190	0.6905	0.6429	0.667	0.5714	0.6905	0.8810	0.6905
F1 Score	0.8513	0.8161	0.8276	0.8371	0.8694	0.8487	0.6727	0.8615
AUC	0.6998	0.7349	0.7238	0.7436	0.7097	0.7525	0.7327	0.7838

Cuadro 3.1: Se muestran diferentes índices para evaluar la calidad de los distintos modelos de clasificación obtenidos a partir del método LDA.

términos generales, superiores en rendimiento a sus contrapartes que no hacen uso de dichas variables. En varios casos sí que se aprecia un diferencia en favor de lo esperable, siendo un claro ejemplo de ello el caso de los modelos QDA, (figura 3.1b). Sin embargo en la mayoría de casos las comparaciones resultan en curvas muy parejas. En los modelos LDA por ejemplo hay mucha proximidad, tanto en los modelos en los que se aplicó la transformación ILR como en los que no.

En cuanto al segundo criterio, a priori, se podría suponer que al comparar un modelo donde se ha realizado la clasificación partiendo de un subconjunto de las variables disponibles, tendría una menor calidad que uno basados en los mismos métodos pero que simplemente usara todas las variables a su disposición. Gráficamente estas comparativas corresponderían a estudiar negro frente a azul y a rojo frente a naranja. En este punto no hay tanta uniformidad en los resultados como en los cruce que estudiamos anteriormente. En figura 3.1a Si comparamos NT-GP con NT-GPV son muy parejos, sin embargo al comparar ILR-GP con ILR-GPV vemos que la curva de este último modelo domina al primero en la mayoría de su dominio. Sin embargo en el caso de RL es claro que NT-GPV es superado por NT-GP así como, con menos contundencia, ILR-GPVdemo supera a ILR-GPdemo. Aunque, siguiendo en RL, el modelo ILR-GP mejora con creces al ILR-GPV a pesar de que NT-GP y NT-GPV se mantienen muy similares, a diferencia de lo que ocurre con RF (figura 3.3), donde NT-GP presenta una mejor clasificación (en términos de curvas ROC) que NT-GPV. De nuevo ILR-GPVdemo supera con significativa diferencia a ILR-GPdemo. Como decíamos previamente, las comparativas en este nivel no presentan un patrón tan claro como en los niveles anteriores.

Aunque las curvas ROC son una herramienta muy útil para hacerse una idea del comportamiento de un clasificador, presentamos a continuación una serie de tablas donde, para cada método de clasificación, LDA, QDA,RL, RF y NB se presentan distintas medidas habituales a la hora de determinar la calidad de un método de clasificación. Cada una de estas medidas se centra en distintos aspectos como ya se detallo en el apartado 2.5. Concretamente el AUC nos servirá para cuantificar y por tanto entender de manera más sencilla las diferencias en las curvas ROC que comentamos anteriormente. En la tabla 3.1 destaca que en todos los casos, al comparar un modelo donde no se ha aplicado la transformación composicional con su equivalente donde sí se ha aplicado se da un incremento entorno al 4 % del valor del AUC



(a) Curvas ROC de los modelos LDA

(b) Curvas ROC de los modelos QDA

Figura 3.1: En ambas imágenes se muestras las curvas ROC de varios modelos, correspondiendo cada imagen a un algoritmo de clasificación distinto. Las curvas punteadas corresponden a modelos donde no se aplicó ninguna transformación y las curvas continuas a modelos donde se aplico la transformación ILR. Los colores codifican las variables empleadas de la siguiente manera: Negro cuando son usadas todas las variables sin incluir las demográficas. Rojo cuando se incluyen las variables demográficas. Azul cuando solo se emplean las variables seleccionadas por VSURF. Naranja cuando se añaden las variables demográficas a las añadidas por VSURF

	Modelos QDA			
Criterio	GPV	ILRV	GPVdemo	ILRVdemo
Exactitud	0.8519	0.8321	0.7660	0.7627
Precisión	0.9800	0.9856	0.9819	0.9843
Sensibilidad	0.8654	0.8396	0.7736	0.7682
Especificidad	0.3810	0.5714	0.5000	0.5714
F1 Score	0.9191	0.9068	0.8654	0.8629
AUC	0.5576	0.7478	0.6768	0.7028

Cuadro 3.2: Se muestran diferentes índices para evaluar la calidad de los distintos modelos de clasificación obtenidos a partir del método QDA.

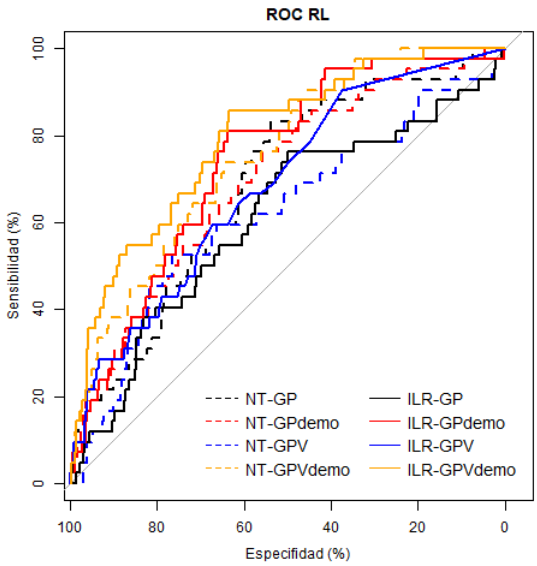


Figura 3.2: Curvas ROC de los modelos de regresión logística. Las curvas punteadas corresponden a modelos donde no se aplicó ninguna transformación y las curvas continuas a modelos donde se aplico la transformación ILR. Los colores codifican las variables empleadas de la siguiente manera: Negro cuando son usadas todas las variables sin incluir las demográficas. Rojo cuando se incluyen las variables demográficas. Azul cuando solo se emplean las variables seleccionadas por VSURF. Naranja cuando se añaden las variables demográficas a las añadidas por VSURF

	Modelos RF							
Criterio	NT-GP	ILR-GP	NT-GPdemo	ILR-GPdemo	NT-GPV	ILR-GPV	NT-GPVdemo	ILR-GPVsocio
Exactitud	0.6213	0.6087	0.4038	0.5036	0.2809	0.3906	0.3364	0.5803
Precisión	0.9849	0.9835	0.9914	0.9865	0.9948	0.9928	0.9937	0.9906
Sensibilidad	0.6200	0.6077	0.3902	0.4963	0.2617	0.3759	0.3195	0.5738
Especificidad	0.6667	0.6429	0.8810	0.7619	0.9524	0.9048	0.9286	0.8095
F1 Score	0.7610	0.7513	0.5600	0.6603	0.4144	0.5454	0.4835	0.7266
AUC	0.6442	0.6333	0.6466	0.6470	0.6113	0.6861	0.6372	0.7145

Cuadro 3.3: Se muestran diferentes índices para evaluar la calidad de los distintos modelos de clasificación RF obtenidos.

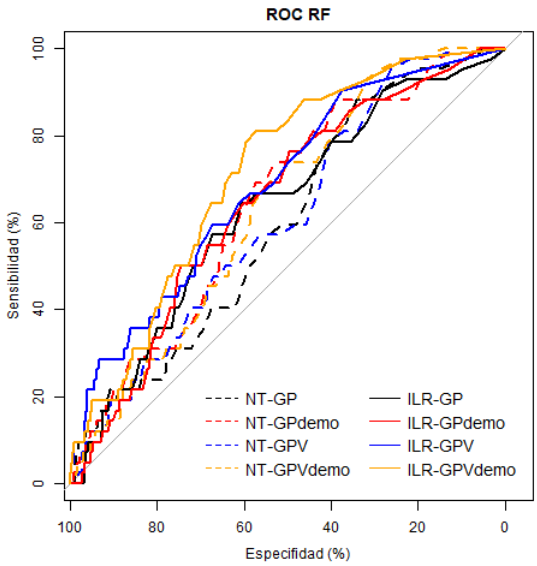


Figura 3.3: Curvas ROC de los modelos de árboles aleatorios. Las curvas punteadas corresponden a modelos donde no se aplicó ninguna transformación y las curvas continuas a modelos donde se aplico la transformación ILR. Los colores codifican las variables empleadas de la siguiente manera: Negro cuando son usadas todas las variables sin incluir las demográficas. Rojo cuando se incluyen las variables demográficas. Azul cuando solo se emplean las variables seleccionadas por VSURF. Naranja cuando se añaden las variables demográficas a las añadidas por VSURF

	Modelos RL							
Criterio	NT-GP	ILR-GP	NT-GPdemo	ILR-GPdemo	NT-GPV	ILR-GPV	NT-GPVdemo	ILR-GPVsocio
Exactitud	0.6682	0.7006	0.8196	0.4600	0.4323	0.8070	0.5208	0.6404
Precisión	0.9850	0.9885	0.9823	0.9970	0.9935	0.9868	0.9934	0.9936
Sensibilidad	0.6689	0.7002	0.8294	0.4460	0.4188	0.8124	0.5105	0.6343
Especificidad	0.6429	0.7143	0.4762	0.9524	0.9048	0.6190	0.8810	0.8571
F1 Score	0.7968	0.8197	0.8994	0.6163	0.5892	0.8911	0.6744	0.7743
AUC	0.6815	0.7493	0.7100	0.7608	0.6996	0.7467	0.7243	0.7930

Cuadro 3.4: Se muestran diferentes índices para evaluar la calidad de los distintos modelos de clasificación RL obtenidos.

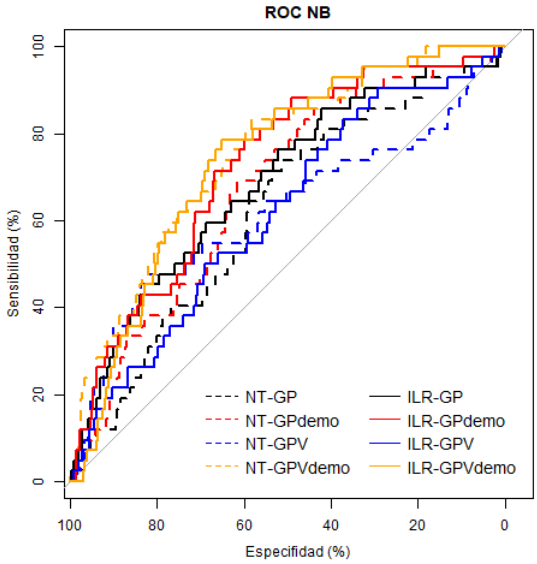


Figura 3.4: Curvas ROC de los modelos Naive Bayes. Las curvas punteadas corresponden a modelos donde no se aplicó ninguna transformación y las curvas continuas a modelos donde se aplico la transformación ILR. Los colores codifican las variables empleadas de la siguiente manera: Negro cuando son usadas todas las variables sin incluir las demográficas. Rojo cuando se incluyen las variables demográficas. Azul cuando solo se emplean las variables seleccionadas por VSURF. Naranja cuando se añaden las variables demográficas a las añadidas por VSURF

	Modelos NB							
Criterio	NT-GP	ILR-GP	NT-GPdemo	ILR-GPdemo	NT-GPV	ILR-GPV	NT-GPVdemo	ILR-GPVsocio
Exactitud	0.5149	0.6841	0.6186	0.6028	0.5023	0.6365	0.4792	0.6557
Precisión	0.9855	0.9834	0.9859	0.9899	0.9891	0.9842	0.9928	0.9907
Sensibilidad	0.5085	0.6866	0.6166	0.5976	0.4935	0.6363	0.4677	0.6519
Especificidad	0.7381	0.5952	0.6905	0.7857	0.8095	0.6429	0.8810	0.7857
F1 Score	0.6709	0.8086	0.7587	0.7452	0.6585	0.7729	0.6359	0.7864
AUC	0.6167	0.6834	0.6676	0.7214	0.6330	0.6588	0.6868	0.7398

Cuadro 3.5: Se muestran diferentes índices para evaluar la calidad de los distintos modelos de clasificación NB obtenidos.

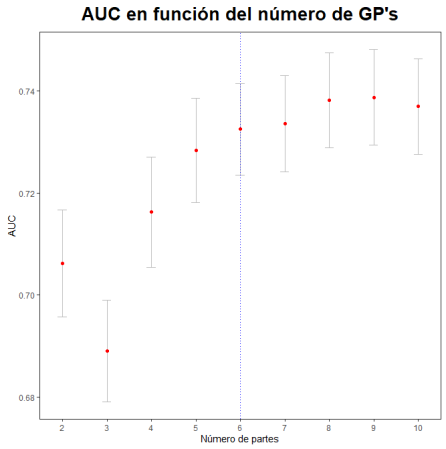


Figura 3.5: En el eje X se presenta el número de partes utilizadas en el balance y en el eje Y el valor de AUC obtenido tras realizar validación cruzada

3.2. Algortimo *selbal*

Una vez explicado el funcionamiento de este algoritmo, utilicémoslo sobre nuestros datos. Este algoritmo primero calcula el balance "óptimo"sobre el total de los datos. Esto hace que el AUC aparente del modelo resultante sea una sobreestimación optimista. Para obtener una medida más realista de la efectividad del modelo se recurre a la validación cruzada con 5 plegamientos por iteración y 10 iteraciones en total. Aún así el balance global puede seguir siendo útil ya que podemos utilizarlo como referencia para comparar con los modelos obtenidos en la validación cruzada y hacernos una idea de cuán robusto es el modelo.

La tabla 3.7 se interpreta de la siguiente manera. La primera columna representa el porcentaje de veces que fue seleccionado un GP como parte del balance en la validación cruzada. Las siguientes columnas indican los GP utilizados en el balance global y en los 3 balances más comunes en la validación cruzada respectivamente, para los cuales se incluye en la última fila el porcentaje de veces que el correspondiente balance fue tomado tras realizar la validación cruzada. Las celdas en rojo indican que el GP se toma en el numerador, en azul en el denominador y en blanco cuando no fue seleccionado. El balance global obtenido sería $\frac{GP_{16}GP_{17}GP_{23}GP_{37}}{GP_{22}}$. Además el balance global parece relativamente robusto pues los GP_{16} y GP_{22} se repiten en los 3 balances más populares en la validación cruzada y el resto de partes se repiten en 2 de los 3.

El AUC obtenido utilizando el balance global es 0.785, en 3.8 se muestra la curva ROC. Como explicamos anteriormente, este balance global utiliza todos los datos y al hacer la predicción sobre los propios datos de entreno produce una estimación excesivamente optimista. Si nos basamos en los AUC calculados en la validación cruzada obtenemos una media de 0.7325, un empeoramiento del 5 %.

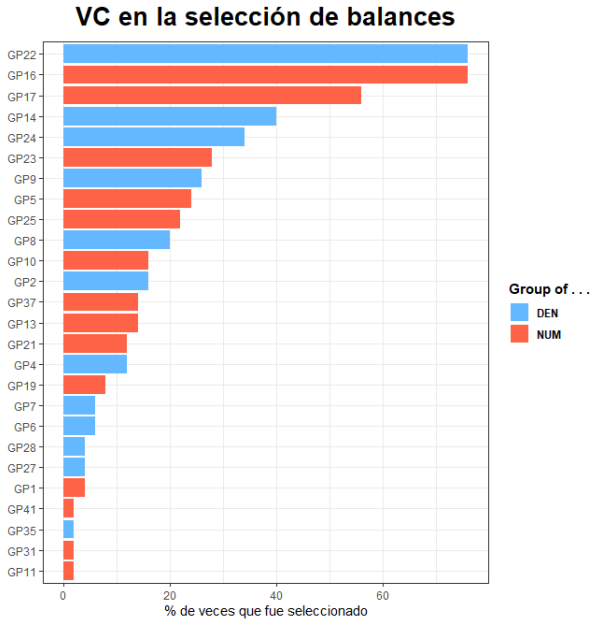


Figura 3.6: En el eje X se presenta el % de veces que un GP fue seleccionado en un balance, y el color representa si fue seleccionado en el numerador (naranja) o en el denominador (azul) del mismo.

Tabla frecuencias de balances

	%	Global	BAL 1	BAL 2	BAL 3
GP16	76				
GP22	76				
GP17	56				
GP14	40				
GP24	34				
GP23	28				
GP25	22				
GP37	14				
FREQ	-	-	0.12	0.1	0.1

Figura 3.7: Representa el balance global obtenido y las partes que lo conforman así como los 3 balances óptimos más veces obtenidos en la validación cruzada. En naranja si la parte pertenece al numerador, en azul si pertenece al denominador y en blanco si no se utilizó en el balance

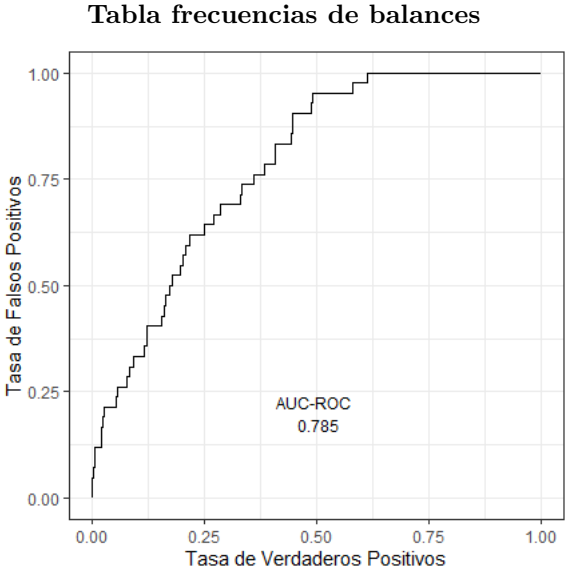


Figura 3.8: Se presenta la curva ROC del balance global obtenido

Capítulo 4

Discusión y conclusiones

4.1. Limitaciones

Una de las limitaciones más importante que afrontamos a lo largo del desarrollo de este trabajo fue el más que notable desbalance entre las clases de la variable respuesta que queríamos estudiar. A pesar de que los algoritmos de clasificación clásicos que se han estudiado incluyen medidas para afrontar este inconveniente, ya fuera pudiendo asignar pesos diferentes a cada individuo, o asignando probabilidades a priori para cada clase de manera personalizada, no encontramos que fueran suficiente, es más, comparándolo con modelos calculados sin aplicar dichas medidas, no se apreciaban mejoras notables. Pese a que desde un principio el encontrar el mejor modelo no fuera el objetivo principal, creemos que este desbalance puede haber tenido un peso en el rendimiento de los modelos que haya minimizado las diferencias que se hallaron entre los modelos donde no se aplicó ninguna transformación composicional y entre los que sí. Además, una parte de estas transformaciones que no pudo explorarse fue la adaptabilidad de las mismas a conocimiento a priori. Como se vio en 2.2.4, concretamente de la transformación SBP, es especialmente útil cuando se tiene algún tipo de noción acerca de posibles agrupaciones de las partes, conocimiento del que no disponíamos en esta ocasión.

Otra limitación fue la falta de adaptación de algunos modelos y metodologías al contexto de datos composicionales, principalmente métodos del análisis funcional. Aunque hay desarrollos teóricos al respecto. (Machalová et al., 2021) no fuimos capaces de adaptarlos e implementarlos correctamente a nuestro estudio. A la hora de interpretar y expresar resultados de manera visual también encontramos limitaciones debidas a la alta dimensionalidad de los datos unida al contexto del análisis composicional, donde la interpretación de las variables de manera individual no es una opción viable pues la información se encuentra en la relación entre variables. Se probó a amalgamar los datos de los N-glucanos para reducir el número de partes pero no encontramos ningún resultado satisfactorio.

Por último, el análisis de datos composicionales en alta dimensión es una rama de la investigación estadística de desarrollo muy reciente, si bien esto resulta motivador, también supone una limitación a la hora de encontrar bibliografía e información práctica acerca de implementaciones prácticas de algunos métodos. Por ejemplo, en análisis de datos funcionales podemos encontrar multitud de información a cerca de distintas distancias, su implementación y análisis a través de simulaciones y casos reales de los distintos efectos que logran cada una de estas medidas. Una variedad tan en datos composicionales (aún) no se da.

4.2. Conclusiones

En este texto se ha dado una visión general de los datos composicionales, una descripción de varias situaciones explicando porqué y cuándo puede llegar a ser útil acudir a este enfoque. Además, la estructura matemática que hay detrás de ellos (recordemos que llega a ser espacio de Hilbert) es suficientemente rica como para pensar que aún quedan muchos métodos por desarrollar, empezando por adaptar métodos ya utilizados en otros contextos, igual que en este trabajo se ha mostrado como se pueden adaptar fácilmente métodos de clasificación clásicos. Al llevar a cabo esta adaptación de los métodos clásicos vemos que, por un lado, la implementación es relativamente sencilla, pues la transformación ILR se logra mediante la función *ilr()* del paquete van den Boogaart et al., 2023 y como se vio en 3 el aplicar métodos composicionales sí que suponen una mejoría. Sin embargo la falta de métodos desarrollados específicamente para datos composicionales pueden haber supuesto que los métodos de clasificación estudiados en este trabajo estén lejos de ser los óptimos.

4.3. Posibles futuras vías de investigación

Creemos que son tres los puntos principales de investigación a partir de este trabajo. La primera sería explorar y estudiar los propios N-glucanos y su estructura, posibles categorizaciones de los GPs podrían conducir a un mejor entendimiento de los mismos y por tanto cuáles son los procesos biológicos en los que intervienen que realmente tienen una relación con el desarrollo del cáncer. Además estas agrupaciones podrían conducir a una reducción de la dimensionalidad de los datos lo cual ayudaría al procesamiento, interpretación y visualización de resultados.

La siguiente vía sería estudiar a fondo el análisis composicional funcional, resultados como los vistos en Talská et al., 2021 son prometedores y creemos que por un lado son el paso lógico a seguir, pues los perfiles de N-glucanos son claramente de ser susceptibles de ser tratados como datos funcionales. Este punto seguramente sería el más interesante a seguir.

En último lugar, pensamos como posible camino a seguir continuar con el desarrollo teórico de herramientas matemáticas en el contexto composicional. Si bien al análisis funcional composicional parece prometedor, creemos que investigar métodos que se mantengan en el mundo composicional, similar a como hace el algoritmo *selbal* trabajando mediante balances puede dar lugar a mejores métodos de clasificación.

Bibliografía

Adamczyk, B., Tharmalingam, T., & Rudd, P. M. (2012). Glycans as cancer biomarkers. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1820*(9), 1347-1353.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139-160.

Billheimer, D., Guttorp, P., & Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American statistical Association*, *96*(456), 1205-1214.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical research*, *65*(12), 4185-4193.

Cheng, K., Zhou, Y., & Neelamegham, S. (2017). DrawGlycan-SNFG: a robust tool to render glycans and glycopeptides with fragmentation information. *Glycobiology*, *27*(3), 200-205.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical geology*, *35*(3), 279-300.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19-20.

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, *51*(4), 1-28. Consultado el 14 de julio de 2023, desde https://www.jstatsoft.org/v51/i04/

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International journal of cancer*.

Gebrehiwot, A. G., Melka, D. S., Kassaye, Y. M., Rehan, I. F., Rangappa, S., Hinou, H., Kamiyama, T., & Nishimura, S.-I. (2018). Healthy human serum N-glycan profiling reveals the influence of ethnic variation on the identified cancer-relevant glycan biomarkers. *PLoS ONE*, *13*(12), e0209515.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2022). *VSURF: Variable Selection Using Random Forests* [R package version 1.2.0]. https://CRAN.R-project.org/package=VSURF

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, *8*, 2224.

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and its Application*, *8*, 271-299.

Gude, F. (2015). *Inflammation and Glycation in a General Adult Population* (Clinical trial registration NCT01796184) [submitted: February 14, 2013]. clinicaltrials.gov. Consultado el 12 de julio de 2023, desde https://clinicaltrials.gov/study/NCT01796184

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357-359.

INE. (2023). Defunciones según la causa de muerte 2022. Datos provisionales.

Lauc, G., Pezer, M., Rudan, I., & Campbell, H. (2016). Mechanisms of disease: The human N-glycome. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1860*(8), 1574-1582.

Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, *10*, 453-472.

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. *R Journal*, *6*(1), 82-92.

Machalová, J., Talská, R., Hron, K., & Gába, A. (2021). Compositional splines for representation of density functions. *Computational Statistics*, *36*(2), 1031-1064.

Mollarasouli, F., Bakirhan, N. K., & Ozkan, S. A. (2022). Chapter 1 - Introduction to biomarkers. En S. A. Ozkan, N. K. Bakirhan & F. Mollarasouli (Eds.), *The Detection of Biomarkers* (pp. 1-22). Academic Press.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, *49*(1/2), 65-82.

Neelamegham, S., Aoki-Kinoshita, K., Bolton, E., Frank, M., Lisacek, F., Lütteke, T., O'Boyle, N., Packer, N. H., Stanley, P., Toukach, P., Varki, A., Woods, R. J., & The SNFG Discussion Group. (2019). Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology*, *29*(9), 620-624.

O'Flaherty, R., Simon, Á., Alonso-Sampedro, M., Sánchez-Batán, S., Fernández-Merino, C., Gude, F., Saldova, R., & González-Quintela, A. (2022). Changes in Serum N-Glycome for Risk Drinkers: A Comparison with Standard Markers for Alcohol Abuse in Men and Women. *Biomolecules*, *12*(2), 241.

Pawlowsky-Glahn, V., & Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, *15*, 384-398.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, *60*(359-367), 489-498.

Rivera-Pinto, J., Egozcue, J. J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., & Calle, M. L. (2018). Balances: a new perspective for microbiome analysis. *MSystems*, *3*(4), e00053-18.

Royle, L., Campbell, M. P., Radcliffe, C. M., White, D. M., Harvey, D. J., Abrahams, J. L., Kim, Y.-G., Henry, G. W., Shadick, N. A., Weinblatt, M. E., Lee, D. M., Rudd, P. M., & Dwek, R. A. (2008). HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Analytical Biochemistry*, *376*(1), 1-12.

Saldova, R., Asadi Shehni, A., Haakensen, V. D., Steinfeld, I., Hilliard, M., Kifer, I., Helland, Å., Yakhini, Z., Børresen-Dale, A.-L., & Rudd, P. M. (2014). Association of N-Glycosylation with Breast Carcinoma and Systemic Features Using High-Resolution Quantitative UPLC [Publisher: American Chemical Society]. *Journal of Proteome Research*, *13*(5), 2314-2327.

Sarmanov, O., & Vistelius, A. (1959). Correlation between percentage variables. *Doklady Akademii Nauk SSSR*, *126*(1), 22-25.

Stöckmann, H., O'Flaherty, R., Adamczyk, B., Saldova, R., & Rudd, P. M. (2015). Automated, high-throughput serum glycoprofiling platform. *Integrative Biology*, *7*(9), 1026-1032.

Talská, R., Hron, K., & Grygar, T. M. (2021). Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*, *53*(7), 1667-1695.

Templ, M. (2021). Artificial neural networks to impute rounded zeros in compositional data. En *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn* (pp. 163-187). Springer.

Van Den Boogaart, K. G., & Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer.

van den Boogaart, K. G., Tolosana-Delgado, R., & Bren, M. (2023). *compositions: Compositional Data Analysis* [R package version 2.0-6]. https://CRAN.R-project.org/package=compositions

Índice de figuras

Índice de cuadros