UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Selección de variables en PLS-DA para la predicción de enfermedad con datos de microbioma

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Beltran Lastra, Samuel

Tutor/a: Tarazona Campos, Sonia

CURSO ACADÉMICO: 2023/2024

Resumen

Entender la relación entre la microbiota y la salud humana puede desempeñar un papel fundamental en la prevención, diagnóstico y tratamiento de las enfermedades. Sin embargo, el análisis estadístico de este tipo de datos conlleva importantes desafíos, ya que se trata de datos de alta dimensionalidad (más variables que observaciones), de naturaleza composicional y con posibles sesgos introducidos por la propia tecnología de secuenciación.

Este trabajo se centra en la aplicación y comparación de modelos de clasificación para la predicción de enfermedades a partir de datos de microbioma, y especialmente en las técnicas de selección de variables, con el objetivo de identificar las bacterias más relacionadas con una determinada enfermedad que puedan ser candidatas a biomarcadores para el diagnóstico o tratamiento de la misma. Específicamente, se utilizará el modelo de clasificación PLS-DA (Regresión en Mínimos Cuadrados Parciales Discriminante) y se compararán diversas estrategias de selección de variables, como el método *Variable Importance in Projection* (VIP), la significación de los coeficientes de regresión, el *Selectivity Ratio* (SR), entre otros. Además, se comparará el rendimiento de las técnicas PLS-DA con respecto a Random Forest y sus técnicas de selección, para evaluar el desempeño de estas frente a uno de los modelos más utilizados en este contexto.

Para ello, se cuenta con 6 bases de datos de microbioma que estudian distintas patologías, como la Diabetes tipo 2, la Cirrosis, entre otras. En cada una de ellas, se dispone de información sobre la microbiota de individuos sanos y enfermos. Se aplicarán las técnicas de selección, se buscarán los hiperparámetros óptimos mediante validación cruzada para optimizar el indicador F1-score, y se compararán los modelos utilizando modelos lineales mixtos.

Finalmente, se interpretarán algunos de los modelos con selección de variables, para analizar si las especies seleccionadas coinciden con los hallazgos reportados por otros estudios.

Palabras clave: PLS-DA; Selección de variables; Microbioma; Modelos de clasificación

Abstract

Understanding the relationship between the microbiome and human health could be crucial for the prevention, diagnosis, and treatment of diseases. However, analyzing this type of data is challenging due to its high-dimensional structure, compositional nature, and biases introduced by sequencing technologies.

This study focuses on implementing and comparing classification models to predict diseases using microbiome data, emphasizing feature selection techniques. The main goal is to identify which species are more closely associated with a particular disease, as these species could serve as potential biomarkers for diagnosis and treatment. Feature selection will allow for the development of more interpretable and accurate models that can be used in a clinical context. Partial Least-Squares Discriminant Analysis (PLS-DA) is the primary model applied, and its performance will be compared with various feature selection techniques such as Variable Importance in Projection (VIP), regression coefficients (RC), Selectivity Ratio (SR), among others. Additionally, PLS-DA techniques will be benchmarked against Random Forest and its feature selection methods, as it is one of the most commonly used models in this type of analysis.

The data consist of six metagenomic datasets spanning diseases such as Type 2 Diabetes, Liver Cirrhosis, among others. Each dataset contains microbiome abundance from both healthy and diseased samples. Classification models with feature selection techniques will be applied to this data. Hyperparameter tuning is performed using a cross-validation structure, maximizing the F1-score, and optimal models are compared using linear mixed models.

Finally, some models with feature selection are interpreted to analyze whether the identified species align with findings from other studies.

Keywords: PLS-DA; Feature selection; Microbiome; Classification models

Agradecimientos

Agradezco a Dios por darme fuerzas, sabiduría y la oportunidad de estudiar en una ciudad como Valencia. Asimismo, quiero expresar mi gratitud a todas las personas que me apoyaron en el desarrollo de este TFM y, en general, del máster. En primer lugar, agradezco a mi tutora del TFM, Sonia, por guiarme en este proceso y brindarme ayuda en cada aspecto que necesité; eres un ejemplo de profesional para mí. En segundo lugar, quiero agradecer a mis padres y a mi hermano, quienes siempre me motivan y apoyan en cada objetivo que me propongo. Finalmente, agradezco a mis amigos del máster por su apoyo y compañía durante todo el proceso académico.

Índice

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1. Microbioma

El microbioma es un conjunto de microbios que se alojan en diferentes partes del cuerpo humano. Estas comunidades se componen de una variedad de microorganismos dentro de los que se incluyen hongos, arqueas, bacterias y virus [1].

En los últimos años, el estudio de la microbiota ha suscitado el interés de la comunidad científica debido al rol que desempeña esta sobre la salud humana, participando en distintas funciones esenciales del cuerpo humano como la digestión, la protección frente a agentes patógenos o procesos metabólicos, entre otros. Por ende, los cambios o alteraciones de estas comunidades influyen de manera determinante en la salud de las personas.

Enfermedades como la obesidad [2], la diabetes [3], la cirrosis [4], la enfermedad inflamatoria intestinal (*Inflammatory bowel disease*, IBD) [5] y la aterosclerosis [6] son algunas de las enfermedades relacionadas a la microbiota tal como se ve en la Figura 1.1.

Figura 1.1: Enfermedades relacionadas a la microbiota [7].

Por otra parte, el estudio de la microbiota y su relación con las enfermedades se ha visto impulsado por el avance de las tecnologías de secuenciación de ADN las cuales han beneficiado la realización de estudios. Dentro de las tecnologías de secuenciación más usadas para la caracterización del microbioma se encuentran las técnicas *"amplicon"* y *"shotgun"*.

La técnica *"amplicon"* se basa en la secuenciación de un marcador genético después de la amplificación por reacción en cadena de la polimerasa. Para bacterias y arqueas, el marcador usado comúnmente debido a su relación costo-efectividad es el gen 16S rRNA ya que contiene las regiones estables y las hipervariables permitiendo clasificar las bacterias y arqueas de la muestra [8].

En el caso del método *"shotgun"* este proceso consiste en fragmentar aleatoriamente una molécula de ADN obteniendo fragmentos que son secuenciados por separado [9]. Esto, a diferencia del método *"amplicon"*, proporciona datos de genes distintos al 16S y permite obtener información de las rutas metabólicas presentes en esa muestra [8].

Así, para ambas tecnologías, la abundancia de un microorganismo (por ejemplo, una bacteria) en la muestra biológica estudiada se estima mediante el número de lecturas de secuenciación obtenidas de dicho microorganismo.

Ambas tecnologías han sido ampliamente usadas en diversos estudios que han permitido llevar a cabo diversos análisis, generalizando y haciendo más consistentes

los patrones encontrados con anterioridad.

En este trabajo de investigación se utilizarán datos de un estudio que recolectó bases de datos de microbioma de distintas patologías, obtenidas mediante secuenciación *shotgun* [10].

1.2. Técnicas estadísticas para el análisis del microbioma

Desde el punto de vista estadístico, el estudio del microbioma ha planteado importantes desafíos. Además del obstáculo de analizar datos de conteos con sobredispersión y abundancia de ceros, se suman la alta dimensionalidad, la naturaleza composicional y los posibles sesgos introducidos por las tecnologías de secuenciación.

Aquí, recopilamos algunas de las principales técnicas estadísticas utilizadas en el análisis de datos de microbiota y sus objetivos.

Preprocesamiento

Uno de los pasos importantes para obtener conclusiones coherentes del análisis es realizar un preprocesamiento adecuado de los datos. Como anteriormente se comentó, las tecnologías de secuenciación pueden introducir sesgos en los datos y por ende, hay que utilizar técnicas de normalización para corregirlos.

Por otra parte, la limitación del número de lecturas de secuenciación por muestra induce a que el aumento de conteos de una bacteria influya directamente en el número de conteos de otra, haciendo que la información de las tablas sea relativa y dependiente de la profundidad de secuenciación. Por esto último, se dice que los datos son composicionales y por ende se tendrán que aplicar técnicas que permitan trabajar con este tipo de datos.

Las técnicas de normalización para la corrección de los sesgos introducidos por la tecnología de secuenciación buscan hacer posible tanto comparaciones entre muestras (*between-samples*) como dentro de una misma muestra (*within-samples*).

Entre las técnicas más utilizadas en el análisis de datos de secuenciación para la corrección de sesgos se encuentra la técnica *Trimmed Mean of M-Values* (TMM) [11], que busca corregir el sesgo *between-samples*. Los pasos consisten en fijar una muestra como referencia, calcular *log-ratios* (*fold changes*) entre el resto de muestras y la referencia, eliminar los valores extremos y los genes más expresados. Luego, se calcula la media del conjunto filtrado para cada muestra, y finalmente, los conteos de

cada muestra son escalados por este factor y por el total de conteos de la muestra. La técnica DESeq [12], que también corrige el sesgo *between-samples*, consiste en calcular la media geométrica de un gen entre las muestras. Luego, para cada muestra, se calcula el ratio entre el conteo y la media geométrica de cada gen. A partir de estos ratios, se obtiene la mediana, y finalmente, los conteos de cada muestra son divididos por esta mediana. Otra normalización es la de Cuantiles [13], que consiste en reemplazar cada cuantil de todas las muestras por la mediana o el promedio de ese cuantil, considerando todas las muestras. La normalización por mediana que consiste en dividir los conteos de la muestra por la mediana de los conteos en esa misma muestra. Por último, RPKM [14] es una técnica que corrige ambos sesgos y se define como el número de lecturas del gen en esa muestra, dividido por la longitud del gen (kb) multiplicado por el número total de lecturas de la muestra en millones.

En el caso de las técnicas de datos composicionales, las transformaciones se basan en el logaritmo de un cierto ratio, esto con el objetivo de eliminar el efecto de abundancia relativa. El precursor del estudio de este campo fue John Aitchison [15] el cual definió las transformaciones *additive log-ratio* (*alr*) y *centered log-ratio transformation* (*clr*) las cuales eliminan el efecto provocado por la dependencia al número total de lecturas y a su vez mantienen la estructura original de correlación subyacente. Es importante tener en cuenta que al eliminar el efecto del total de conteos por muestra no es requisito imprescindible aplicar otra técnica de normalización previa.

Una de las técnicas de preprocesamiento que se aplicaron en los conjuntos de datos que se usarán en este trabajo es la transformación *clr* para datos composicionales, la cual corresponde al logaritmo del ratio entre la abundancia de la especie y la media geométrica de las abundancias de todas las especies, por lo que todas las partes se tratan simétricamente. Además, en este caso se le suma una unidad al numerador, ya que por la naturaleza de los datos de microbioma, estos tienden a tener muchos ceros.

Otro método aplicado en los conjuntos de datos analizados es la normalización *Total Sum Scaling* [16], donde cada abundancia de la especie se divide por la profundidad de secuenciación de la muestra i y a este ratio se le suma una unidad para evitar los ceros para finalmente aplicar el logaritmo. Esta técnica permite eliminar el sesgo asociado a la profundidad de secuenciación y se usa como alternativa al método *clr* en los conjuntos de datos que se analizarán.

Modelos estadísticos

Las técnicas estadísticas aplicadas en el análisis del microbioma son diversas y dependen del objetivo del estudio.

Los modelos no supervisados apuntan a la identificación de patrones y correlaciones entre los microorganismos estudiados. Dentro de estas técnicas podemos encontrar el análisis de componentes principales (*PCA*) [17]. Esta técnica permite explorar tanto relaciones entre pacientes como relaciones entre las bacterias. Otras técnicas similares utilizadas son el análisis de coordenadas principales (*PcoA*) [18], Factorización de matrices no negativas (*NMF*) [19] y *"t-distributed stochastic neighbor embedding (t-SNE)"* [20].

A diferencia de los modelos no supervisados, en los supervisados se define una variable respuesta. Esta variable es la que se quiere predecir a partir de los predictores. En el caso del análisis de datos de microbioma, generalmente la variable respuesta corresponde al indicador de si la muestra corresponde a un individuo sano o enfermo y las variables predictoras corresponden a las especies de bacterias. Por tanto, los modelos a utilizar son de clasificación.

Modelos como regresión logística [21], análisis discriminante (*LDA*) [22], vecinos más cercanos (*KNN*) [23], *Naive Bayes* [24], máquinas de soporte vectorial (*SVM*) [25] y Random Forest (*RF*) [26] son algunos de los modelos de clasificación implementados en los últimos estudios [27]. La aplicación de estos modelos por sí solos se ve dificultada por el reducido tamaño muestral y la alta cantidad de especies de bacterias, por lo que el uso de estos modelos usualmente va acompañado de técnicas de selección de variables.

Aunque estos modelos suelen tener una alta capacidad predictiva, no son tan buenos a la hora de seleccionar los predictores más influyentes sobre la variable respuesta para la mejor interpretación del modelo, especialmente en escenarios de alta multicolinealidad y/o dimensionalidad, como es el caso del microbioma. Un modelo alternativo de clasificación es el de regresión en mínimos cuadrados parciales en su versión discriminante (PLS-DA) [28]. Esta es una técnica estadística muy útil en el análisis de datos de microbioma, ya que no solo permite ajustar un modelo de predicción, si no que otorga una gran interpretabilidad, ya que se puede utilizar en datos de alta dimensionalidad y se beneficia de la multicolinealidad de los predictores, mediante la definición de variables latentes que son combinación de los predictores originales y describen los patrones de comportamiento de los predictores que más influyen en la variable respuesta.

En un trabajo anterior [29], se comparó el modelo PLS-DA con los modelos Random Forest (RF) y Support Vector Machine (SVM) en seis bases de datos públicas de microbioma procedentes del estudio de Pasolli et al. [10], a las que se les aplicó distintos métodos de preproceso. Sin embargo, en dicho trabajo, no se evaluó ninguna estrategia

de selección de variables. El presente trabajo se centra, pues, en evaluar distintas técnicas de selección de variables en PLS-DA y, a su vez, compararlas con algunas técnicas de selección de variables en RF, modelo que ha mostrado alta capacidad predictiva en este tipo de estudios.

1.2.1. Selección de variables

La optimización de un modelo predictivo busca que este se ajuste de la mejor manera a los datos con los que se está trabajando. A la vez, trata de que el modelo no se sobreajuste a los datos de entrenamiento y sea parsimonioso. Además, dependiendo del estudio también se puede requerir que el modelo sea interpretable.

En el caso de los datos de microbioma se posee una gran cantidad de variables predictoras (normalmente cientos de bacterias) y los estudios realizados se centran en la búsqueda de subconjuntos de bacterias que permitan un predicción óptima de la enfermedad o condición. A este conjunto de variables predictoras se les denomina biomarcadores. Por esto, las técnicas de selección de variables se tornan un aspecto fundamental en estos estudios, debido a que pueden mejorar tanto la capacidad predictiva de los modelos como la interpretabilidad de estos, además de que los biomarcadores pueden tener utilidad clínica en el tratamiento y/o diagnóstico de la enfermedad.

Existen diversos métodos de selección de variables, que se pueden clasificar en tres categorías principales: *"filter"*, *"wrapper"* y *"embedded"*.

Las técnicas *"filter"* consisten en calcular una métrica o puntaje que mida la relevancia de las variables predictoras con respecto a la variable respuesta. Luego, se ordenan las variables según esta métrica, se filtran definiendo un umbral o *"threshold"*, y así se obtiene el conjunto de variables seleccionadas. Algunos ejemplos de técnicas de tipo *"filter"* son los test univariados, como ANOVA, en los que se realiza un análisis de varianza de cada variable predictora con respecto a la variable respuesta. Otros ejemplos incluyen *limma* [30] y *sam* [31], que también son test univariados con ciertas variaciones y son muy utilizados en el análisis de expresión génica. Además, la técnica de importancia de variables por permutación del Random Forest también entra en esta clasificación [32]. *Minimum Redundancy Maximum Relevance* (mRMR) [33] selecciona los predictores con mayor *"mutual information"* (MI) respecto a la variable respuesta, pero con la menor MI posible respecto a los predictores seleccionados previamente. Una de las ventajas de estos métodos es que son independientes del modelo de clasificación, es decir, el conjunto de variables seleccionado se puede usar como entrada en el modelo

de clasificación finalmente evaluado. Además, computacionalmente son menos costosos que los métodos *"wrapper"* y *"embedded"*. En el caso del PLS, *"Variable Importance in Projection"* (VIP) [34] y coeficientes de regresión (*"Regression coefficients"*, RC) [35] son de las más utilizadas y se utilizan en este trabajo.

Los métodos *"wrapper"* realizan la selección del conjunto óptimo de variables en combinación con el modelo. Es decir, se realiza la evaluación de los distintos subconjuntos de variables respecto al impacto que tienen sobre la precisión de un modelo específico. Por ejemplo, los métodos *stepwise forward and backward* [36] son de este tipo, ya que, se van añadiendo o eliminando variables de acuerdo al criterio seleccionado. la significación del coeficiente, el criterio de información Akaike (AIC) y el criterio de información de Bayes (BIC) son algunos de los criterios utilizados. Algoritmo Genético (GA) [37], *"Statistically Equivalent Signature"* (SES) [38] y Boruta [32] son otras técnicas *"wrapper"* aplicadas en análisis de datos de expresión génica. Los métodos *"wrapper"* son más costosos computacionalmente, ya que la búsqueda del conjunto de variables óptimo se realiza de manera iterativa en búsqueda del mejor rendimiento del modelo.

La última categoría *"embedded"* se refiere a técnicas donde la selección de variables está integrada en el ajuste del modelo. Al estar integradas en el modelo, su contexto de aplicación se acota a ese tipo de modelo en particular. Las más utilizadas son las técnicas de selección de regularización como *least absolute shrinkage and selection operator* ($LASSO$) [39]. Esta técnica incorpora en el ajuste del modelo una restricción o penalización sobre los coeficientes de regresión, haciendo que algunos coeficientes se contraigan a cero eliminándolos del modelo. Dependiendo de la elección del parámetro de penalización se filtrarán más o menos variables predictoras. Otra técnica de regularización es *Elastic Net* [40]. Esta técnica incorpora dos restricciones de penalización, una de tipo $LASSO$ y otra de tipo *Ridge* [41]. Esto lo hace ser una técnica más flexible y permite superar dos limitaciones importantes de $LASSO$. La primera, es que en el caso que $p > n$, $LASSO$ seleccionaría como máximo n variables y esto es una desventaja sobre todo en los datos de microbioma. Segundo, con grupos de variables altamente correlacionadas podría seleccionar solo una de estas variables y desechar el resto de las variables del grupo, aunque estas sean significativas.

Esta investigación se centra en la comparación de técnicas de selección de variables para el modelo PLS-DA. Estas técnicas se aplicarán a los datos públicos utilizados en [10] y procesados en la tesis de máster "Análisis de datos de microbioma para la predicción y caracterización de enfermedades" [29]. En concreto, se evaluará la capacidad predictiva del modelo PLS-DA tras la selección de variables y su interpretabilidad, así

como si la estrategia de preproceso aplicada a los datos influye en estos resultados.

Capítulo 2

Objetivos

2.1. Objetivo General

En el marco del análisis de datos de microbioma, uno de los principales objetivos es encontrar relaciones entre las especies de microorganismos y la enfermedad en estudio. En estudios recientes se ha corroborado, mediante la aplicación de técnicas estadísticas, la relación que existe entre la microbiota y ciertas enfermedades. Gran parte de estas técnicas se enfocan en la predicción y no permiten una óptima interpretabilidad de la relación entre las especies y la enfermedad. Por esto, el modelo PLS surge como una técnica estadística muy potente en este contexto, ya que por un lado permite obtener un modelo de predicción y, a la vez ofrece herramientas que facilitan la detección de patrones entre las especies y la enfermedad o condición.

Aunque, esta técnica es muy eficiente en contextos de alta dimensionalidad y ruido, complementarla con técnicas de selección de variables suele mejorar tanto su precisión como interpretabilidad. En esa línea, este trabajo se enfoca en la aplicación y comparación de técnicas de selección de variables del modelo PLS-DA aplicado a datos de microbioma, pudiendo así obtener modelos más interpretables y detectar el conjunto de especies que más influyen en la ausencia o presencia de la enfermedad.

2.2. Objetivos Específicos

- Revisión bibliográfica para la elección de técnicas de selección de variables para PLS y PLS-DA.
- Aplicación de dichas técnicas de selección de variables a cada una de las bases de datos del estudio preprocesadas mediante distintas estrategias.

- Comparación y evaluación del desempeño de las técnicas de selección de variables en PLS-DA de acuerdo con su capacidad de predicción y estabilidad.
- Comparación del modelo PLS-DA con sus técnicas de selección respecto al modelo Random Forest, con y sin selección de variables.
- Analizar la relación de las variables seleccionadas por los distintos métodos y ver cómo influyen en la presencia o ausencia de la enfermedad.

Capítulo 3

Materiales y Métodos

3.1. Datos

Las bases de datos utilizadas en este trabajo provienen de una recopilación realizada por Pasolli et al. [10], trabajo en el que comparan técnicas de machine learning sobre conjuntos de datos provenientes de varios estudios. Los datos obtenidos mediante técnicas de secuenciación de tipo *"shotgun"* estudió cinco patologías. Cirrosis, cáncer colorrectal (CRC), enfermedad inflamatoria intestinal (IBD), la obesidad y diabetes tipo 2 (WT2D y T2D).

En estas bases de datos las variables corresponden a la abundancia relativa de las bacterias y las filas son los sujetos incluidos en el estudio (casos y controles). Además, en el estudio original se incluyen otras covariables, como variables demográficas que en este estudio no se utilizaron.

A partir de estos conjuntos de datos, se aplicaron en un trabajo de máster previo [29] diversas estrategias de preprocesamiento con el fin de mejorar la capacidad predictiva de los modelos de clasificación aplicados y de reducir los sesgos introducidos por la tecnología de secuenciación.

La primera etapa de preprocesamiento corresponde a un filtrado previo de los datos con el fin de eliminar bacterias con baja abundancia para así eliminar el ruido y mejorar el análisis estadístico.

Se realizaron dos tipos de filtro.

- Filtro S1 : Se eliminan las especies de bacterias con abundancia cero en todas las muestras del conjunto.
- Filtro S2 : Se utilizarán las variables (especies) que se encuentran presentes en más de un 20 % de las muestras.

Tabla 3.1: Dimensiones de los conjuntos de datos

Enfermedad	# Variables S1	# Variables S2	# Casos	# Controles
Cirrosis	530	190	118	114
Cáncer Colorrectal (CRC)	493	184	48	73
Diabetes Tipo 2 (T2D)	556	166	170	174
Diabetes Tipo 2 (WT2D)	367	162	52	43
Enfermedad Inflamatoria Intestinal (IBD)	425	175	25	85
Obesidad	446	155	164	89

Las dimensiones de los conjuntos de datos una vez aplicados los filtros se observan en la tabla 3.1. Por otra parte, se aplicaron estrategias de normalización para eliminar posibles sesgos que se generan debido a la tecnología de secuenciación.

Para explicarlas, introduciremos la notación del problema. Sea $\boldsymbol{X}$ $(N \times K)$ nuestra matriz de datos, donde N corresponde al número de individuos del estudio y K son las columnas que en este caso corresponden a las especie de bacterias. Cada elemento de la matriz x_{ij} representará la abundancia de la especie j en el individuo i donde $i = 1, \ldots, N$ y $j = 1, \ldots, K$. Las normalizaciones que se aplicaron fueron dos:

- Total Sum Scaling (TSS): En esta normalización cada abundancia de la especie se divide por la profundidad de secuenciación de la muestra i y a este ratio se le suma una unidad para evitar los ceros y se aplica el logaritmo [16]. Esta técnica elimina el sesgo asociado a la profundidad de secuenciación y en el trabajo se denominó como normalización N1. Además, en este caso el ratio se multiplicará por 10^6 para evitar trabajar con números muy pequeños.

$$TSS_{ij} = log(\frac{x_{ij}}{S_i}10^6 + 1) \tag{3.1}$$

- Aitchison Centered Log Ratio (CLR): Esta normalización proviene del campo del análisis de datos composicionales y fue propuesta por Aitchison [15]. Corresponde al logaritmo de un ratio donde el numerador es la abundancia de la especie j en el individuo i y el denominador es la media geométrica G_i de todos los elementos. Esto hace que todos los elementos se traten de manera simétrica. Para evitar los ceros en el logaritmo se suma una unidad a todos los elementos de la matriz de datos. En el trabajo corresponde a la normalización N2.

$$CLR_{ij} = log(\frac{x_{ij} + 1}{G_i}) \tag{3.2}$$

$$G_i = \sqrt[k]{\prod_{j=1}^{k}(x_{ij} + 1)} \tag{3.3}$$

Concluyendo, en el trabajo se analizaron 24 conjuntos de datos que se generaron a partir de la aplicación de los filtros y normalizaciones descritos, tal como se ve en la figura 3.1.

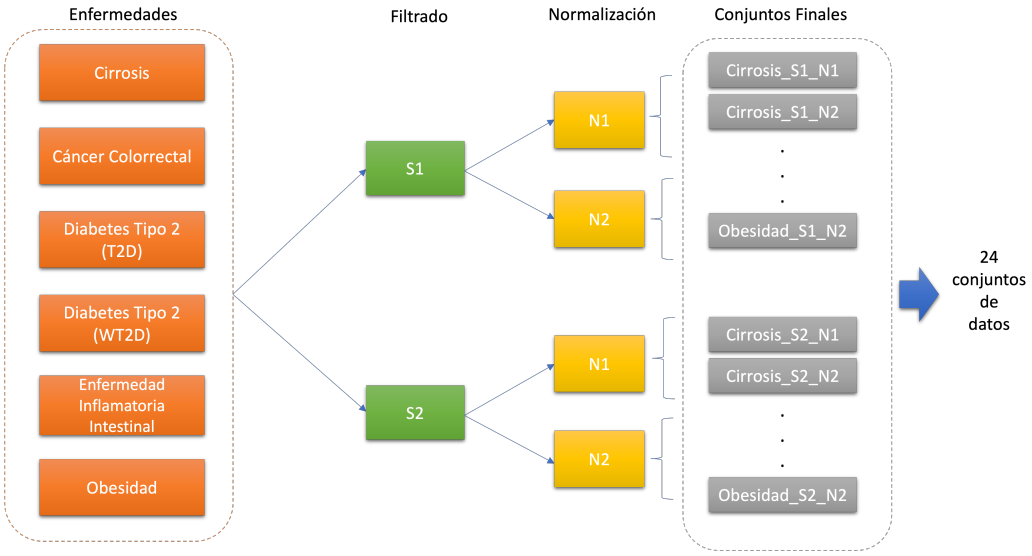

Figura 3.1: Conjuntos de datos a utilizar

3.2. Metodología

3.2.1. PLS

Una de las técnicas más utilizadas y estudiadas cuando se quiere predecir una variable respuesta numérica a partir de diversas variables predictoras es el modelo de regresión lineal múltiple (MLR). Este modelo posee ciertas limitaciones y no se comporta adecuadamente en entornos donde se posee un gran número de predictores, altamente correlacionados y con un alto ruido. Además, no permite que el número de predictores sea mayor al número de observaciones. Por otra parte, en el caso de existir múltiples variables respuesta se necesitará ajustar un modelo para cada una de ellas.

En este contexto, Wold desarrolló el modelo de regresión en mínimos cuadrados parciales (PLS) [34]. Esta técnica estadística multivariante busca maximizar la covarianza entre la matriz de predictores $\boldsymbol{X}$ y la matriz de respuestas $\boldsymbol{Y}$, a la vez que reduce la dimensionalidad de ambas matrices explicando la mayor cantidad de variabilidad de las mismas.

Al ajustar el modelo PLS sobre un conjunto de datos de N observaciones, K predictores y M variables respuesta, la descomposición de las matrices $\boldsymbol{X}$ $(N \times K)$ e $\boldsymbol{Y}$ $(N \times M)$ queda expresada de la siguiente manera:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E} \tag{3.4}$$

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{C}^{\boldsymbol{T}} + \boldsymbol{F} \tag{3.5}$$

Donde $\boldsymbol{X}$ se descompone como una matriz de scores $\boldsymbol{T}$ multiplicada por la matriz de loadings transpuesta más la matriz de residuos $\boldsymbol{E}$.

En el caso de $\boldsymbol{Y}$ se expresa como la matriz de scores $\boldsymbol{T}$ multiplicada por la matriz de pesos $\boldsymbol{C}$ transpuesta más la matriz de residuos $\boldsymbol{F}$.

Si observamos la descomposición de la matriz $\boldsymbol{X}$, realizando una analogía con un modelo de regresión, los loadings son los "coeficientes de regresión" que indican la relación entre los scores y $\boldsymbol{X}$. Por ende, los loadings se pueden interpretar como la contribución de cada variable de $\boldsymbol{X}$ en los scores.

Por otra parte, los scores t_a (donde $a = 1, 2, ..., A$. Siendo A el número de componentes) también se pueden definir como una combinación lineal entre las variables x_k y los pesos w^*_{ka}, además estos vectores son ortogonales entre sí. En forma matricial se expresa como:

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}^* = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-\mathbf{1}} \tag{3.6}$$

Donde $\boldsymbol{W}$ es la matriz de pesos. Los pesos w son las correlaciones de las variables de $\boldsymbol{X}$ con u (scores de Y). Por ende, los scores son la combinación lineal entre la matriz $\boldsymbol{X}$ y la matriz de pesos $\boldsymbol{W}^*$. De manera interpretativa, esto implica que los scores capturan la variabilidad de $\boldsymbol{X}$ que se correlaciona con la variabilidad de $\boldsymbol{Y}$. Si reemplazamos esta última expresión en $\boldsymbol{Y}$ obtenemos:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-\mathbf{1}}\boldsymbol{C}^{\boldsymbol{T}} + \boldsymbol{F} \tag{3.7}$$

Por tanto, el modelo PLS se puede representar como un modelo de regresión:

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{B}}_{\boldsymbol{PLS}} + \boldsymbol{F} \tag{3.8}$$

Donde la matriz de coeficientes del modelo se calcula como:

$$\hat{\boldsymbol{B}}_{\boldsymbol{PLS}} = \boldsymbol{W}(\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W})^{-\mathbf{1}}\boldsymbol{C}^{\boldsymbol{T}} \tag{3.9}$$

Es importante señalar que el número de componentes A del modelo PLS es un parámetro a definir. Un número pequeño de componentes podría no capturar toda la variabilidad de los datos y con un número alto el modelo podría sobre ajustarse

a los datos de entrenamiento y/o incluir ruido que no aporta al modelo. Por esto, para obtener un balance entre bondad de ajuste y predicción en el modelo, se utiliza validación cruzada para determinar el número óptimo de componentes.

3.2.1.1. PLS Discriminante (PLS-DA)

Del modelo PLS deriva su variante el PLS Discriminante (PLS-DA), modelo que se utiliza cuando la variable respuesta es categórica. En este caso, el modelo pasa a ser un modelo de clasificación.

Al ser una extensión, se mantiene la estructura del modelo PLS pero, la variable respuesta, que en este caso es categórica, se transforma en tantas variables dummies (0-1) como categorías existan.

El modelo PLS-DA en este trabajo tiene la estructura de la figura 3.2 ya que la variable respuesta en este caso indica si el individuo posee o no la enfermedad.

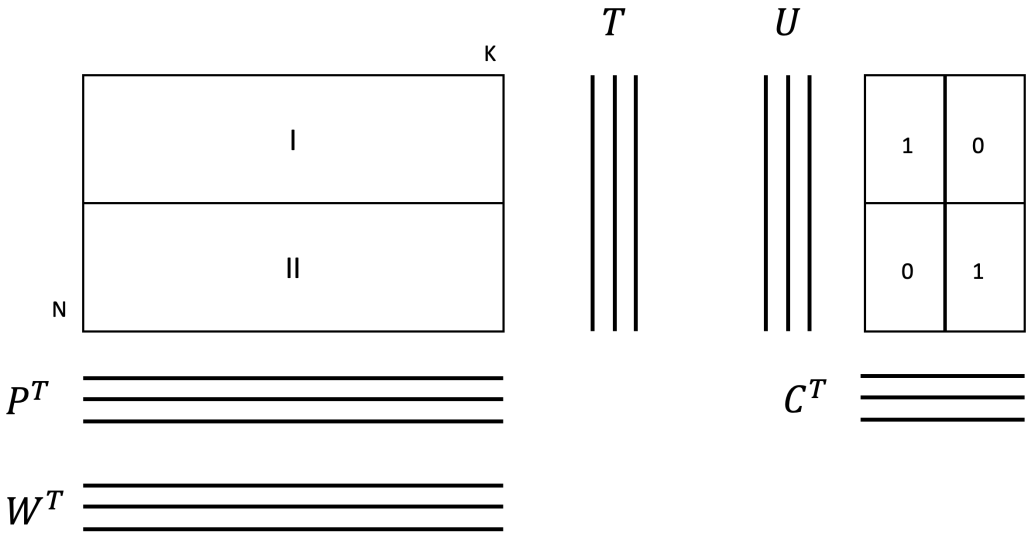

Figura 3.2: Diagrama PLS-DA

3.2.2. Selección de Variables en PLS-DA

Aunque el modelo PLS-DA se desenvuelve adecuadamente en conjuntos de datos de alta dimensionalidad la inclusión de variables poco relevantes puede influir negativamente en la capacidad predictiva del modelo. Además, la complejidad del modelo aumenta y por ende la interpretabilidad de este se dificulta. Por esto, la selección de variables puede ayudar en estos entornos ya sea eliminando variables que no aporten capacidad predictiva al modelo como también permitiendo una mejor

interpretabilidad.

Existen una diversidad de técnicas de selección de variables en el PLS. En este estudio se realizó una revisión sistemática de estas y se seleccionaron las que mejor rendimiento obtuvieron en investigaciones anteriores. Luego, estas se implementarán en el modelo PLS-DA con el objetivo de mejorar su capacidad predictiva y su interpretabilidad.

Como se verá a continuación, las técnicas de selección de variables tienen hiperparámetros que deben ser definidos. En este trabajo, la búsqueda de los hiperparámetros óptimos se realiza mediante validación cruzada k-fold repetida, con $k = 10$ folds y $R = 10$ repeticiones. Finalmente, se elige la combinación de valores de los hiperparámetros que entrega el mejor F1-score medio.

Las técnicas de selección de variables que se aplican en este estudio se pueden clasificar en tres categorías: *"filter"*, *"wrapper"* y *"embedded"* [42], como se ha comentado en la introducción.

3.2.2.1. Métodos *"Filter"*

Las técnicas de selección *"filter"* o de filtrado se componen de dos fases. Primero, se ajusta el modelo PLS con todas las variables, luego se selecciona una medida de importancia de las variables calculada a partir del modelo. Las variables que estén por sobre o debajo de cierto umbral de esa medida serán seleccionadas o filtradas. Si se selecciona un umbral fijo, las técnicas *"filter"* se pueden aplicar como una técnica de selección independiente al modelo de clasificación finalmente ajustado.

El principal desafío de este tipo de métodos es la elección del umbral de la medida de importancia, ya que este límite será el que determine si una variable es seleccionada o no.

En este trabajo, la búsqueda de ese umbral óptimo se realiza mediante la optimización de los hiperparámetros del modelo. Esto implica que el método de selección está inserto en la validación cruzada, transformando las técnicas *"filter"* en *"wrapper"*, ya que los conjuntos de variables seleccionados se evaluarán en un modelo de clasificación específico.

Algunas de las medidas que cuantifican la importancia de la variable predictora en el modelo PLS son: los coeficientes de regresión (*"Regression coefficients"*, RC), "*Variable Importance on Projection*" (VIP) y *"Selectivity Ratio"* (SR).

Coeficientes de Regresión (RC)

Los coeficientes de regresión son una medida de la relación entre una variable predictora y la variable respuesta.

Del modelo PLS también se pueden extraer los coeficientes de regresión de cada uno de las predictores obteniendo un valor de su impacto sobre la variable dependiente.

Para la selección de variables mediante este método podría establecerse un valor umbral h, así cuando un β esta por debajo del valor h se filtra y en caso contrario se mantiene.

Otro enfoque desarrollado y que se aplica en este trabajo consta del cálculo de la significación de los coeficientes mediante un test de permutaciones [35]. El primer paso, consiste en generar una distribución aleatoria del coeficiente de regresión bajo la $H_0 : \beta_j = 0$. Luego, se compara el valor del coeficiente real con los valores aleatorios (distribución nula). Si el valor aleatorio es mayor al real, en términos absolutos, quiere decir que la variable no es estadísticamente significativa.

El algoritmo en detalle consta de los siguientes pasos:

1. Ajustar el modelo PLS-DA sobre las matrices $\boldsymbol{X}$ $(N \times K)$ e $\boldsymbol{Y}$ $(N \times M)$.

2. Se obtienen los coeficientes de regresión del modelo PLS-DA , $\hat{\boldsymbol{B}}_{\boldsymbol{PLS}}$ $(K \times M)$.

3. Se genera una distribución aleatoria permutando las observaciones de la variable respuesta, en este caso, $\boldsymbol{Y}$ se permuta R veces. Para cada permutación se ajusta un modelo PLS-DA, obteniendo R modelos PLS-DA. Finalmente, se tienen R coeficientes de regresión para cada predictor.

4. Calculamos el p-valor de cada predictor j con la siguiente fórmula:

$$\text{Pvalor}_j = \frac{\sum_{g=1}^{R} A_g}{R} \tag{3.10}$$

$$\text{donde } A_g = \begin{cases} 1 & \text{Si } |\beta_g| > |\beta_{j-real}| \\ 0 & \text{Si } |\beta_g| \leq |\beta_{j-real}| \end{cases}$$

$j = 1, ..., K$

$R =$ número de permutaciones

$\beta_g =$ coeficiente de regresión aleatorio de la iteración g

$\beta_{j-real} =$ coeficiente de regresión real de la especie j

5. Se ordenan los p-valores de cada especie de menor a mayor, es decir, desde los más significativos a los menos significativos y se elige una cantidad H de variables a seleccionar.

6. Con las H variables más significativas seleccionadas se ajusta un nuevo PLS-DA con el cual se predice finalmente.

El valor H de variables a seleccionar es el hiperparámetro a optimizar en la búsqueda del modelo óptimo.

Variable Importance in Projection (VIP) El método *"variable importance in projection"* (VIP) es una medida de la importancia de la variable predictora en el espacio de proyección. En este caso la importancia de la variable j se define como:

$$VIP_j = \sqrt{\frac{p \sum_{a=1}^{A} [c_a^2 t_a^T t_a (w_{aj}/\|w_a\|^2)]}{\sum_{a=1}^{A} c_a^2 t_a^T t_a}} \tag{3.11}$$

$A =$ Número total de componentes del modelo

$p = K =$ Total de variables

$w_{aj}/\|w_a\|^2 =$ Importancia de la variable j en la componente a

$c_a^2 t_a^T t_a =$ Varianza de y explicada por la componente a

El procedimiento que se aplica en este trabajo consta de los siguientes pasos:

1. Ajustar el modelo PLS-DA.
2. Calcular el VIP para cada variable j.
3. Seleccionar las variables con un VIP $> h$, donde h es el umbral de selección.
4. Finalmente, se ajusta el modelo con las variables seleccionadas.

El valor h es el límite de corte establecido para el VIP. Comúnmente, se utiliza el criterio de seleccionar a las variables con un VIP > 1 como significativas aunque, para este caso se buscará el valor óptimo del VIP en la optimización de los hiperparámetros iterando 30 valores VIP para cada una de las bases de datos estudiadas.

Selectivity Ratio (SR)

Al ajustar un modelo PLS-DA, este puede requerir de varias componentes para lograr una buena separación de las clases. Esto conlleva a que la interpretación del

modelo se vuelva más compleja. Así nace este método, que se basa en aplicar un procedimiento llamado *"Target Projection"* que tiene por objetivo simplificar el modelo PLS-DA, generando un "modelo" con una sola componente. Esto lo realiza proyectando la descomposición de $\boldsymbol{X}$ (obtenida del PLS-DA) sobre la variable respuesta $\boldsymbol{Y}$ [43]. Así, el modelo se puede expresar como:

$$\boldsymbol{X} = \hat{\boldsymbol{X}}_{\boldsymbol{TP}} + \boldsymbol{E}_{\boldsymbol{TP}} = \boldsymbol{t}_{\boldsymbol{TP}}\boldsymbol{p}_{\boldsymbol{TP}}^{\boldsymbol{T}} + \boldsymbol{E}_{\boldsymbol{TP}} \tag{3.12}$$

Donde $t_{TP} = \boldsymbol{X}\frac{\hat{\boldsymbol{B}}_{\boldsymbol{PLS}}}{\|\hat{\boldsymbol{B}}_{\boldsymbol{PLS}}\|}$ y $p_{TP} = \boldsymbol{X}^T\frac{t_{TP}}{t_{TP}^T t_{TP}}$ [44]. Los loadings de este nuevo modelo serían los "coeficientes" que sirven para medir cuánto es la contribución de una variable predictora en el modelo PLS. Con esto, se define la medida Selectivity Ratio (SR) como:

$$SR_j = \frac{v_{exp,j}}{v_{res,j}} = \frac{SCE_j}{SCR_j} = \frac{\sum_{i=1}^{N}\hat{x}_{TP_{ij}}^2}{\sum_{i=1}^{N}e_{TP_{ij}}^2} \tag{3.13}$$

Considerando que $v_{exp,j}$ y $v_{res,j}$ es la varianza explicada y residual de la variable j donde $j = 1, \ldots, K$. Mayores valores de SR explican una mayor capacidad de discriminación. Para la selección de un límite de SR, el artículo plantea varias opciones como una prueba F y una prueba no paramétrica [43]. Sin embargo, en este trabajo se optimizará el valor de corte SR mediante la validación cruzada. Por esto, en esta implementación se iterarán 60 valores de SR optimizando este hiperparámetro con el fin de obtener las variables que optimicen la capacidad predictiva de los modelos para cada una de las bases de datos.

3.2.2.2. Métodos *"Wrapper"*

Tal como se mencionó en la introducción, los métodos *"wrapper"* son aquellos que realizan la selección del conjunto óptimo de variables con respecto a un modelo de clasificación específico. Es decir, al buscar el conjunto óptimo de variables, se generan y evalúan distintos subconjuntos respecto a un modelo en particular con el fin de optimizar su precisión. Esto implica que los modelos *"wrapper"* son iterativos, ya que se prueban diversos subconjuntos, lo que implica mayor riesgo de sobreajuste y también que sean computacionalmente más costosas que las técnicas de filtrado [45].

Algunas de las técnicas *"wrapper"* de selección en PLS son *"Montecarlo variable elimination"* (MVE), *"Genetic Algorithm"* con PLS (GA-PLS), *"Iterative predictor weighting PLS"* (IPW-PLS), *"Backward variable elimination"* (BVE-PLS), T^2

Hotelling, entre otras [46].

Las que se aplicaron en este trabajo son BVE-PLS y T^2 Hotelling luego de que obtuvieran los mejores resultados en el estudio comparativo [46].

Backward Variable Elimination (BVE-PLS)

Para aplicar esta técnica es necesario seleccionar una medida de importancia de las variables. En este trabajo utilizamos el valor VIP. Esto porque el método consiste en filtrar las variables poco relevantes eliminando aquellas que estén por debajo de un valor establecido.

Luego, se vuelve ajustar el modelo y se realiza nuevamente este proceso de filtrado. A cada modelo ajustado se puede medir su capacidad predictiva de manera que al terminar el proceso iterativo se selecciona el modelo con el mejor rendimiento.

Para este caso, se creó una nueva función basada en el algoritmo BVE-PLS implementado en la librería *plsVarSel* [42]. Los pasos del algoritmo implementado son los siguientes:

1. Ajustar el modelo PLS-DA con los datos de entrenamiento.
2. Calcular el vector de predicciones usando el conjunto de test.
3. Calcular el VIP de las variables.
4. Eliminar variables cuyo VIP esté por debajo del límite h.
5. Actualizar los conjuntos de entrenamiento y test con las variables seleccionadas en el paso anterior.
6. Volver al paso 1 e iterar hasta que no se puedan filtrar más variables, ya sea porque el número de variables es menor o igual al número de componentes o, porque ya no hay más variables con un VIP por debajo del límite h para filtrar.

Como se observa, la principal diferencia con la técnica de selección VIP del apartado *"filter"* es que en este caso se itera con las variables que van quedando seleccionadas y se ajusta un nuevo modelo.

Un aspecto a considerar en la implementación realizada es que no en todos los folds se iterará la misma cantidad de veces. Esto variará dependiendo de las muestras que estén incluidas en ese fold, por lo que finalmente se considerará el número de iteraciones máximo que haya ocurrido en todos los folds para esa repetición de la validación cruzada.

T^2 Hotelling

Este método de selección se basa en los pesos w del modelo PLS y en el estadístico T^2 de Hotelling para la selección de las variables informativas. Se asume, por el teorema central del límite que los pesos w se distribuyen normalmente. Extendiendo esta idea, la matriz de pesos $\boldsymbol{W}$ seguirá una distribución normal multivariante. El enfoque del método es construir un intervalo de confianza para el T^2 y así diferenciar los valores de los pesos que se diferencian del resto (ruido) [47]. Los pasos de este método son:

1. Se extrae la matriz de pesos $\boldsymbol{W}$ del modelo PLS ajustado con todo el conjunto de variables.
2. Se calcula el estadístico T^2 a partir de esa matriz de pesos $\boldsymbol{W}$.:

$$T^2 = n(\bar{W}_i - \bar{\bar{W}})^T S_W^{-1}(\bar{W}_i - \bar{\bar{W}}) \tag{3.14}$$

n = Tamaño de muestra

$i = 1, ..., K$ = Número de variables

$\bar{W}_i$ = Media de pesos de la observación i

$\bar{\bar{W}}$ = Media de la media de los pesos

S_W^{-1} = Inversa de matriz de varianzas covarianzas de pesos

Al iterar i, el tamaño de muestra n se convierte en 1 y $\bar{W}_i$ son los pesos de la observación i. Finalmente, se obtendrán tantos valores de T^2 como número de variables K existan.

3. Se establece el límite superior (UCL) que determinará las variables informativas. Este corresponde a:

$$UCL = C(p, A^*)F_{(A^*, p-A^*, \alpha_{T^2})} \tag{3.15}$$

$$C(p, A^*) = \frac{A^*(p-1)}{p-A} \tag{3.16}$$

A^* = Número de componentes del modelo PLS-DA

p = Número de variables = K

α_{T^2} = Nivel de significación para límite de T^2

4. Las variables que tengan un $T^2 > UCL$, es decir, los *outliers* son clasificadas como variables informativas y las demás variables son filtradas.

5. Finalmente, se ajusta el modelo PLS con las variables seleccionadas y se mide el rendimiento del modelo.

α_{T^2} determinará el UCL y por ende el número de variables que se filtrarán. En este caso α_{T^2} será un parámetro a optimizar en la búsqueda del modelo óptimo.

3.2.2.3. Métodos *"Embedded"*

A diferencia de los métodos anteriores en los que la selección estaba fuera del ajuste del modelo PLS, en los métodos *"embedded"* la selección de variables y el ajuste del modelo PLS se realizan simultáneamente.

Algunas de las técnicas de este tipo son *"Interactive variable selection"* (IVS), *"soft-threshold PLS"* (ST-PLS), *"sparse-PLS"* (SPLS), entre otras [42].

En este trabajo, se ha utilizado SPLS, por lo que se describe a continuación.

Sparse-PLS (SPLS)

Para explicar la técnica de selección SPLS hay que expresar el ajuste del modelo PLS como un problema de optimización. El objetivo es encontrar los vectores de pesos w_a (a = 1,..., A). Para el caso donde $\boldsymbol{Y}$ es univariado, los vectores se pueden encontrar resolviendo el siguiente problema:

$$\max_{w} \text{cor}^2(Y, Xw)\text{var}(Xw) \tag{3.17}$$

En el caso del SPLS [48] incorpora la selección de variables planteando el siguiente modelo de optimización:

$$\min_{w,c} -\kappa w^T M_{XY} w + (1-\kappa)(c-w)^T M_{XY}(c-w) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2 \tag{3.18}$$

Donde $w^T w = 1$, $M_{XY} = X^T Y Y^T X$, c son vectores de pesos "auxiliares", w son los vectores de pesos y $\kappa, \lambda_1, \lambda_2$ son parámetros de penalización a optimizar. Como se observa, el problema de optimización cambió al incorporar dos penalizaciones. La primera penalización de tipo L_1, se encarga de imponer una restricción sobre el vector c en vez de directamente sobre el vector w, manteniéndose cercanos entre ellos al mismo tiempo. Por otra parte, la penalización L_2 maneja la posible singularidad de M_{XY}.

La solución a este problema en el caso de Y univariado (0/1) considerando $\lambda_2 = \infty$ nos entrega un vector de pesos "auxiliares":

$$\hat{c} = \left(|Z| - \frac{\lambda_1}{2} \right)_+ \text{sgn}(Z) \tag{3.19}$$

Donde $Z = \frac{X^T Y}{\|X^T Y\|}$ y $(x)_+ = max(0, x)$. Además, hay que considerar que, en el caso univariado, independientemente del valor del parámetro κ, se llega a la misma solución de la ecuación anterior [48].

Esta solución de tipo *"soft thresholding"* se ha reformulado [48] obteniendo:

$$\hat{c} = \left(|Z| - \eta \max_{1 \leq j \leq K} |Z_j| \right)_+ \text{sgn}(Z) \tag{3.20}$$

Ahora, solo se debe optimizar el parámetro η que tomará valores entre 0 y 1. Si es igual a 0, se convierte en un PLS sin selección. Por otro lado, si se aumenta el valor de este parámetro implicará seleccionar una menor cantidad de variables. La búsqueda del valor óptimo de η se realizará en la optimización de los hiperparámetros del modelo.

3.2.3. Random Forest (RF)

Con el objetivo de comparar el modelo PLS-DA y sus diversos métodos de selección de variables, con otros modelos predictivos, se seleccionó el modelo RF por sus buenos resultados en otros estudios comparativos sobre datos del microbioma y porque también permite, al menos, seleccionar los predictores más importantes.

Random Forest es un modelo de bagging, es decir, es un modelo que se compone de diversos clasificadores. En este caso, los modelos son múltiples arboles de decisión que se generan a partir de subconjuntos de datos y variables seleccionados aleatoriamente [26].

El algoritmo del modelo consta de los siguientes pasos:

1. A partir del conjunto de entrenamiento se crean múltiples subconjuntos de datos mediante *bootstrap*.

2. Para cada subconjunto de datos se ajustan árboles de decisión. Una diferencia principal con un árbol de decisión tradicional es que en Random Forest se selecciona un conjunto aleatorio de las variables predictoras en cada nodo y respecto a estas se hace la ramificación. Este procedimiento se repite nodo a nodo hasta el crecimiento máximo del árbol, es decir, no se realiza una poda.

3. Para realizar las predicciones se realiza una agregación de cada predicción individual de cada árbol. En el caso de un problema de clasificación se elige la

clase más votada y en regresión se promedia el resultado de cada árbol.

4. Para la estimación del error del modelo se utilizan las muestras *Out-of-Bag* (OOB). Es decir, todas las muestras que no fueron utilizadas para la generación del subconjunto de datos mediante *bootstrap*. Para cada una de estas muestras se realiza una predicción utilizando los árboles donde la muestra no fue utilizada en el subconjunto de datos. Se agregan las predicciones de las OOB dependiendo de si es un problema de regresión o clasificación y se calcula una métrica de error comparando las predicciones con los valores reales de las OOB. Esto entrega una medida del error del modelo Random Forest.

Respecto a la interpretabilidad del modelo RF, una de las técnicas más utilizadas para medir la importancia de variables es de permutar los valores de una variable predictora y observar la disminución de la precisión del modelo. Si disminuye considerablemente la precisión significa que es una variable importante para el modelo.

Por ejemplo, para calcular la importancia de la variable j :

1. Se permutan los valores de la variable j en las muestras *OOB*.
2. Se calcula el error de predicción del modelo con la variable permutada.
3. La importancia de la variable se puede calcular como:

$$\text{Importancia de la Variable}_j = \text{Error OOB permutada}_j - \text{Error OOB original} \quad (3.21)$$

3.2.4. Selección de Variables en Random Forest

Al igual que en el modelo PLS y en otros modelos estadísticos, la selección de variables es un paso conveniente en el modelo Random Forest para generar un modelo parsimonioso.

Existen múltiples técnicas de selección para aplicar en el modelo Random Forest, sin embargo, no es el objetivo de este trabajo y solo se analizaron tres técnicas de selección que son *Boruta*, *varSelRF* y *VSURF*.

Para el caso de Boruta y varSelRF se obtuvieron resultados. Sin embargo, en la implementación del método *VSURF*, la técnica no seleccionó ninguna variable en varios conjuntos de datos, no pudiendo realizar las predicciones y obtener resultados válidos para esas bases de datos. Por esto, se descartó del estudio. Esto también ocurrió en el artículo comparativo de las técnicas de selección de Random Forest, cuando se probó en conjuntos de datos con alta cantidad de predictores, por lo que *VSURF* se recomienda para conjuntos con poco ruido y con menos de 50 predictores [49].

Boruta

Esta técnica de selección de variables del tipo *wrapper* esta basada en el modelo Random Forest y extiende el concepto del cálculo de la importancia de una variable mediante permutación. Esto lo hace agregando algunos pasos que permiten detectar mejor la importancia de una variable al minimizar el impacto de los efectos aleatorios y la correlación [50]. Los pasos del algoritmo son:

1. Crear el conjunto de datos "extendido". Esto se hará añadiendo copias de las variables pero con los valores permutados. Estas copias permutadas de las variables son denominadas *shadow variables*.

2. Se ajusta un modelo Random Forest sobre el nuevo conjunto de datos extendido y se calcula la importancia de las variables calculando un Z score a partir de la disminución de precisión del modelo.

$$Z = \frac{\text{Disminución promedio precisión}}{\text{Desviación estandar disminución precisión}} \tag{3.22}$$

3. Se determina el Z score máximo de las *shadow variables* (*Maximum Z Score of Shadow Features*,MZSA). Este se puede interpretar como la mayor importancia de un efecto aleatorio.

4. Se comparan los Z scores de las variables originales con el MZSA. Si el Z score de la variable es mayor al MZSA se marca como importante en esa repetición.

5. Para determinar finalmente si una variable es importante o irrelevante se realiza un test bilateral de igualdad con el MZSA.

6. En el caso de que el atributo sea significativamente mayor que el MZSA se clasifica como una variable importante. Por otra parte, si es significativamente menor se clasifica como irrelevante.

7. Se eliminan las variables *shadow*.

8. Este proceso se repite hasta el límite de iteraciones definidas o hasta que todas las variables hayan sido clasificadas como importantes o irrelevantes.

varSelRF

Este método de selección fue desarrollado con el objetivo de evaluar el rendimiento del modelo de Random Forest aplicado a datos de expresión génica medida mediante *microarrays* y ver su capacidad de selección de genes en este tipo de datos [51].

El objetivo de los autores de esta técnica fue encontrar el subconjunto de variables más pequeño que tuviera una buena capacidad predictiva. El fin es poder usar este subconjunto en un contexto clínico y de diagnóstico.

Los pasos del algoritmo son:

1. Se ajusta un modelo Random Forest con todas las variables.
2. Se calcula la importancia de las variables usando la técnica de permutación de los valores de cada variable y midiendo su aporte en la disminución promedio del error del modelo.

$$\text{Importancia de la Variable} = \text{Error OOB variable permutada} - \text{Error OOB variable original} \tag{3.23}$$

3. Se filtran las variables con menor importancia. El algoritmo por defecto filtra un 20 % de las variables menos relevantes.
4. Con el conjunto de variables seleccionadas se vuelve ajustar un modelo de Random Forest y se calcula el error con las muestras OOB .
5. Este proceso se realiza iterativamente repitiendo el paso 3 y 4 hasta que se cumpla una condición de salida. La condición de salida por defecto se produce cuando el error OOB de la iteración actual es mayor al error OOB de la iteración anterior o inicial más una desviación estándar.
6. Finalmente, se selecciona el subconjunto de variables con el error OOB más bajo.

Un aspecto muy importante a destacar de esta técnica de selección es que realiza un filtrado de variables de manera iterativa, lo que provoca que se evite la selección de variables redundantes, es decir, altamente correlacionadas.

Esto se alinea con el objetivo del método que es obtener un conjunto de variables para el diagnóstico. Sin embargo, esto afecta a la estabilidad del modelo ya que distintos conjuntos de variables pueden ser óptimos y obtener una buena capacidad predictiva.

3.2.5. Optimización y Validación de Modelos

Validación Cruzada

Para la validación de los modelos entrenados, se utilizó un procedimiento de validación cruzada. Este tiene por objetivo minimizar el sobreajuste que se provoca al

entrenar un modelo con un conjunto de datos específico, proporcionando una mejor estimación del rendimiento del modelo.

En este caso, se utilizó la validación cruzada k-fold, la cual consiste en dividir el conjunto de datos en k grupos. Luego, se ajusta el modelo utilizando $k-1$ grupos como conjunto de entrenamiento y el grupo restante como conjunto de prueba. Este proceso se repite k veces utilizando en cada iteración un fold distinto como conjunto de prueba [52]. Como resultado de las k iteraciones se obtendrá el vector de predicción. En este trabajo se utilizó $k = 10$.

Además, con el objetivo de añadir otra capa de aleatoriedad, este proceso de ajuste se repetirá R veces. En este trabajo se estableció un $R = 10$. Esto nos permitirá mejorar la estimación de la precisión del modelo ya que finalmente obtendremos 10 vectores de predicción los cuales compararemos con la variable respuesta real.

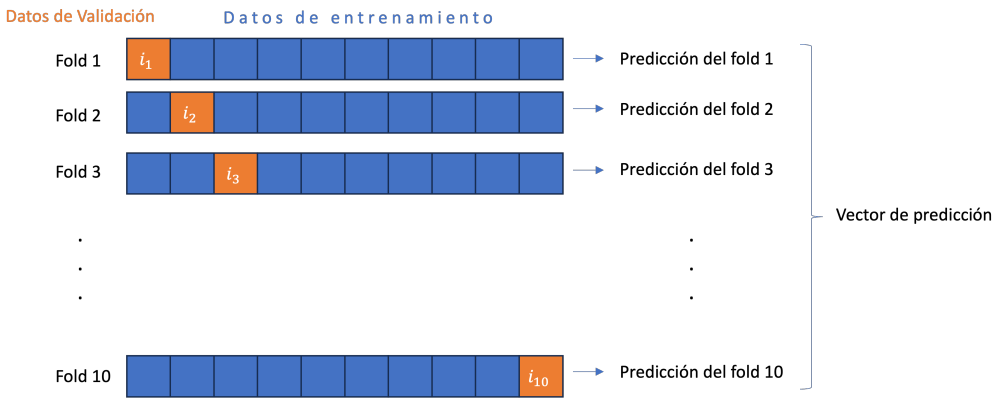

Figura 3.3: Validación Cruzada k-fold

Matriz de Confusión y Evaluación del Modelo

Para evaluar los modelos de clasificación, una de las herramientas utilizadas es la matriz de confusión. Esta permite evaluar el desempeño del modelo de clasificación comparando los valores predichos con los valores reales.

En este caso, la variable respuesta binaria se compone de casos, que son los que poseen la enfermedad, y controles, que son los individuos sanos.

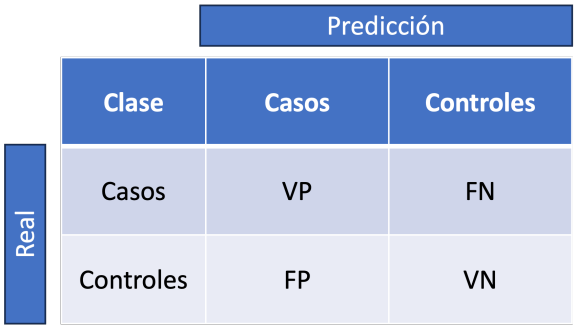

Figura 3.4: Matriz de Confusión

Para definir posteriormente las métricas comúnmente utilizadas, hay que definir los elementos que componen la matriz de confusión:

- Verdaderos Positivos (VP): son los casos que el modelo predijo correctamente como casos.
- Verdaderos Negativos (VN): son los controles que el modelo predijo correctamente como controles.
- Falsos Positivos (FP): son los controles que el modelo predijo incorrectamente como casos.
- Falsos Negativos (FN): son los casos que el modelo predijo incorrectamente como controles.

En este trabajo, se utilizaron las siguientes métricas:

1. *Precision* (Precisión): es la proporción de casos correctamente clasificados como positivos respecto al total de los predichos como positivos.

$$\text{Precision} = \frac{VP}{VP + FP} \tag{3.24}$$

2. *Recall* (Sensibilidad): es la proporción entre los casos correctamente clasificados como positivos respecto al total de casos reales positivos. Es decir, es la capacidad del modelo para detectar los casos positivos (enfermos).

$$\text{Sensibilidad} = \frac{VP}{VP + FN} \tag{3.25}$$

3. F1-Score: es la media armónica de la precisión y la sensibilidad. Al mezclar la precisión y la sensibilidad, provee un balance al integrar métricas que consideran tanto los falsos positivos como los falsos negativos. Esta es la métrica que se

utiliza para comparar los modelos, ya que es particularmente útil en contextos donde las clases no están balanceadas, ya sea que haya más controles que casos o viceversa.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensibilidad}}{\text{Precision} + \text{Sensibilidad}} \tag{3.26}$$

Optimización de hiperparámetros de los modelos

Para todos los modelos ajustados se utilizó la validación cruzada k-fold repetida. Los parámetros utilizados para los diversos modelos se pueden dividir en dos secciones: los generales, que son comunes para todos los modelos ajustados en el estudio, y los específicos, que dependen del modelo y la técnica de selección aplicada. Los parámetros generales son:

- Número de componentes: para los modelos PLS-DA, con y sin selección, se itera de 1 hasta 10 componentes.
- Punto de corte: para la definición de caso o control de las predicciones, se itera el punto de corte desde 0 hasta 1 en intervalos de 0.05.

Por otra parte, están los parámetros específicos para las técnicas de selección del PLS-DA:

- **VIP**: el parámetro a optimizar será el punto de corte VIP. Se iteraron 30 valores para cada conjunto. Los valores van desde 0 hasta el máximo VIP de ese conjunto de datos (el máximo VIP posible se calcula ajustando un PLS con 1 componente para esa base de datos).
- **Coeficientes de regresión**: para calcular la significación de una variable se realizan 1000 permutaciones. Por otra parte, el parámetro a optimizar es el número de variables a retener, este parámetro varía en intervalos de 20 variables, comenzando en 20 hasta el máximo de predictores que tenga la base de datos. Este intervalo podría ser más pequeño, pero aumentaría el costo computacional; por ello, se decidió utilizar el mismo intervalo que en el estudio en el que se basó [35].
- **Selectivity Ratio**: el parámetro a optimizar es el punto de corte SR. Se iteraron 60 valores para cada base de datos. Los valores van desde 0 hasta el máximo SR de esa base (el máximo SR posible se calcula en cada iteración ya que varía mucho dependiendo del número de componentes).

- **SPLS** : el parámetro a optimizar es η. Se iteró entre 0 y 0.9 en intervalos de 0.1.
- $\boldsymbol{T^2}$: el parámetro a optimizar es α. Se iteró entre 0.05 y 1 en intervalos de 0.05.
- **BVE**: el parámetro a optimizar es el punto de corte VIP. Considerar que este filtro VIP se realiza de manera iterativa en esta técnica de selección del tipo *wrapper*.

Para el modelo de Random Forest, se utilizaron los parámetros establecidos por Pasolli [10] $ntree = 500$ y $mtry = \sqrt{p}$. Los parámetros de las técnicas de selección fueron:

- **Boruta** : los parámetros son los que trae la función de la librería *Boruta* por defecto.
- **varSelRF**: de la función de la librería *varSelRF* se utilizaron los parámetros $ntree = 500$, $ntreeIterat = 500$ y $whole.range =$ FALSE, este último con el fin de no expandir totalmente el árbol y usar el error OOB como criterio de parada.

Como las técnicas de selección de variables están construidas en base al Random Forest, para los parámetros que correspondan al RF, principalmente *mtry* y *ntree*, se utilizan los parámetros recomendados en la literatura, donde $mtry = \sqrt{p}$ [53] [54] y $ntree = 500$ [55] [56], considerando que con estos se obtienen buenos resultados y su optimización con escasa frecuencia mejora los modelos. Por ende, para el RF y sus técnicas de selección, el único parámetro a optimizar será el punto de corte de la probabilidad de clasificación, que al igual que en el modelo PLS-DA y sus técnicas de selección, se iterará de 0 a 1 en intervalos de 0.05.

Al realizar la CV, para ciertas combinaciones de hiperparámetros, no se pudieron estimar los modelos y calcular la medida del error para todos los folds de esa repetición. En dichos casos, se optó por descartar la combinación de hiperparámetros que, quizás por ser demasiado restrictiva, no generaba modelos válidos. Tras descartar estas combinaciones de hiperparámetros para cada método y base de datos, se seleccionó la combinación de hiperparámetros que entrega el mayor F1-score medio. En caso de empate, se eligió la que tuviera menor desviación típica del F1-score.

Una vez determinados los hiperparámetros óptimos, se procede al ajuste de los modelos finales utilizando estos valores. Para ello, se emplea la misma estructura de CV k-fold (10 particiones) con 10 repeticiones, pero utilizando una semilla distinta a la empleada en la búsqueda de los hiperparámetros.

Estabilidad

Otro aspecto que se medirá es la estabilidad de las técnicas de selección. La estabilidad se refiere a la consistencia en la selección de las variables al ajustar diversos modelos con distintos conjuntos de entrenamiento.

En términos prácticos, una técnica de selección estable tenderá a seleccionar las mismas variables aunque varíen los conjuntos de entrenamiento. Esta característica es deseable, ya que los predictores se utilizan para interpretar el modelo y también podrían utilizarse en la generación de biomarcadores para la detección de enfermedades, por lo que variaciones grandes podrían conllevar interpretaciones erróneas [57].

Para medir la estabilidad se usó una adaptación de la distancia de Tanimoto que se expresa como:

$$S_S(s, s') = 1 - \frac{|s| + |s'| - 2|s \cap s'|}{|s| + |s'| - |s \cap s'|} \tag{3.27}$$

Esta es una medida de similaridad, que en este caso se calcula entre dos conjuntos de variables seleccionadas s y s', donde $s = (s_1, s_2, ..., s_j), s_i \in \{0, 1\}$ y $j = 1, ..., K($ Total de variables de la base de datos sin selección $)$. Por lo tanto, cada conjunto de variables seleccionadas quedará representado por un vector con valores 0/1, donde 0 significa que la variable no fue seleccionada y 1 que sí lo fue.

Esto nos permite medir la similitud entre dos conjuntos de variables seleccionadas. La adaptación permite una mejor interpretabilidad, ya que cuando dos conjuntos comparados son idénticos, la métrica toma el valor de 1 y cuando no existe ninguna coincidencia entre los dos conjuntos, toma el valor de 0.

Esta métrica se calculó con los modelos finales, es decir, usando los resultados de los modelos ajustados con la combinación de hiperparámetros óptimos.

La estabilidad se calculó agrupada por los factores de pre procesamiento , técnica de selección, enfermedad y repetición. Los pasos son los siguientes:

1. Se calcula la distancia de Tanimoto entre los distintos pares de conjuntos de variables seleccionadas. Como dentro de una repetición se generan 10 conjuntos de variables seleccionadas (provienen de cada uno de los 10 folds), el número de comparaciones será de $\frac{10(10-1)}{2} = 45$, obteniendo así 45 valores de estabilidad.

2. La estabilidad de la repetición será la media de los 45 valores calculados en el paso anterior. Con esto se obtiene una distancia media de Tanimoto para cada repetición del conjunto de datos.

3.2.6. Comparación de modelos y técnicas de selección

Para comparar los modelos y medir los efectos de los distintos factores involucrados se utilizaron modelos lineales mixtos (MLM).

Un MLM se formula de la siguiente manera [58]:

$$y = X\beta + Zb + \varepsilon \tag{3.28}$$

Donde:

y = Vector variable respuesta

X = Matriz de diseño - efectos fijos

β = Vector de coeficientes - efectos fijos

Z = Matriz de diseño - efectos aleatorios

b = Vector de coeficientes - efectos aleatorios

ε = Vector de errores

En este trabajo se ajustaron dos modelos lineales mixtos. El primero, se realizó para ver la influencia de los efectos, fijos y aleatorios, sobre la variable respuesta F1-score y el segundo para ver la influencia de los mismos efectos sobre la estabilidad.

Los efectos fijos simples que se incluyen en los modelos son: técnica de selección (técnica), preprocesamiento y base de datos (BBDD). Además, se incluyen las interacciones dobles de los factores anteriores: técnica-preprocesamiento, técnica-BBDD y preprocesamiento-BBDD.

En el caso de los efectos aleatorios, se incluye el efecto de la repetición en cada base de datos (BBDD:rep). Esto se debe a que, en el proceso de ajuste, se realizaron 10 repeticiones de la CV k-fold para cada combinación de los niveles de los factores, lo que podría tener un impacto en la variabilidad de la variable respuesta.

Luego, para realizar las comparaciones por pares, primero se calculan las medias marginales estimadas (MME). Las MME son las medias de las predicciones del modelo para cada nivel de un factor (o la combinación de varios) ajustadas, considerando los niveles de los otros factores. Para esto se utiliza la librería *emmeans* [59]. Finalmente, se utiliza la prueba de Tukey para realizar las comparaciones de las MME.

3.2.7. Software

Para el análisis, modelado, manipulación de datos, creación de visualizaciones y otras tareas análisis estadístico se utilizó el lenguaje de programación R y diversas librerías que se detallan a continuación:

- *caret* [60]: se utilizaron las funciones de creación de folds y la función para ajustar el PLS-DA (*plsda*).
- *plsVarSel* [61]: esta es una de las librerías principales respecto a las técnicas de selección del modelo PLS. Se utilizaron las funciones de *VIP*, *RC* (coeficientes de regresión) y *SR*. También, sobre algunas funciones de esta librería se basaron las implementaciones realizadas de las técnicas de selección T^2 y *BVE*.
- *spls* [62]: se utilizó esta librería para el ajuste de la técnica spls, utilizando la función *splsda*.
- Librerías de procesamiento paralelo: se utilizaron las librerías *parallelly* [63], *parallel*, *doRNG* [64] y *doParallel* [65] para el procesamiento paralelo de los modelos y técnicas de selección ajustadas con el fin de mejorar los tiempos de ajuste y asegurar la reproducibilidad.
- Random Forest y técnicas de selección: se utilizaron las librerías *randomForest* [66], *Boruta* [32] y *varSelRF* [67].
- Manipulación de datos: se utilizó la librería *tidyverse* [68].
- Modelo lineales mixtos : para ajustar los MLM se utilizó la librería *lme4* [69]. Para el cálculo de las medias marginales estimadas y sus comparaciones por pares se utiliza la librería *emmeans* [59].

Capítulo 4

Resultados

4.1. Optimización de los modelos de clasificación con y sin selección de variables

Como se mencionó en la metodología, para la elección de los hiperparámetros óptimos de cada modelo se aplica dicho modelo con distintos valores de los parámetros, y finalmente, se elige la combinación de hiperparámetros que entrega el mejor F1-score medio. En caso de empate, se elige la de menor desviación típica. Además, para la elección se descartaron aquellas combinaciones de hiperparámetros en las que no se ajustaron modelos satisfactoriamente para las 10 repeticiones de la CV k-fold. Esto se observó en casos donde el hiperparámetro de la técnica era muy restrictivo, es decir, no seleccionaba ninguna variable y por ende no podía ajustar el modelo en el fold y como consecuencia en la repetición.

Hay que considerar que, sobre los 24 conjuntos de datos, se ajustan 10 modelos diferentes, los cuales pueden incluir o no una técnica de selección. Por lo tanto, en total se tienen 240 modelos ajustados.

Para ejemplificar el proceso de elección de la combinación óptima de hiperparámetros, se muestran los resultados del modelo PLS-DA con selección mediante VIP del preprocesamiento S2-N2 para la enfermedad Cirrosis en la Tabla 4.1. Aquí, se observan los 10 mejores resultados del F1-Score medio y sus hiperparámetros respectivos. En este caso, la combinación óptima es cuando el número de componentes es igual a dos, límite VIP es 0.60 y el punto de corte es 0.40. Es decir, usando esta combinación se consigue el mejor de los modelos en términos de F1-score medio. De esta forma se elige la combinación óptima de hiperparámetros para cada técnica, preprocesamiento y base de datos. Por ejemplo, la Tabla 4.2 contiene los resultados de los hiperparámetros

Tabla 4.1: S2-N2 Cirrosis PLS-VIP

Núm. Componentes	VIP	Punto de Corte	Media F1	Desviación estándar F1
2	0.60	0.40	0.827	0.012
2	0.60	0.45	0.824	0.011
2	0.48	0.45	0.824	0.012
2	0.36	0.45	0.823	0.009
2	2.04	0.30	0.821	0.012
2	0.72	0.40	0.821	0.009
2	0.48	0.40	0.820	0.009
2	0.00	0.45	0.820	0.008
2	0.96	0.45	0.820	0.014
2	0.12	0.45	0.819	0.006

óptimos de los modelos del preprocesamiento S2-N2. Esto se realiza para cada modelo, para así usar estos hiperparámetros en el ajuste de los modelos finales con los que se comparan las técnicas de selección. Los resultados de los otros preprocesamientos se encuentran en el anexo A.2.

Para el ajuste de los modelos finales se usa la misma estructura de CV k-fold (10 folds) con 10 repeticiones. Es decir, ajustamos los modelos con los hiperparámetros óptimos utilizando otra semilla aleatoria para comparar los modelos finales.

Aunque, en la selección de hiperparámetros óptimos se habían descartado las combinaciones de valores restrictivas, en el ajuste final hubo un modelo en que la combinación óptima no entregó resultados para todas las repeticiones. Específicamente, ocurrió en la BBDD de Obesidad para el preprocesamiento S1-N1 con el método $PLS - T^2$. Los hiperparámetros utilizados fueron: número de componentes = 5, punto de corte = 0.1 y un $\alpha_{T^2} = 0.05$. El motivo fue que al ajustar el modelo sobre folds generados con la nueva semilla, en algunos folds no se seleccionó ninguna variable porque el α_{T^2} era demasiado restrictivo. El detalle de las repeticiones incompletas se encuentra en el anexo A.3.

4.2. Comparación de modelos y técnicas de selección de variables

Una vez optimizados los distintos modelos y técnicas de selección de variables se procedió a su comparación. Para ello, se consideraron dos aspectos: la métrica F1-score y la estabilidad (Tanimoto). Como se mencionó en la metodología, se ajustaron dos

Tabla 4.2: Parámetros óptimos para normalización S2-N2

Técnica	Parámetro	WT2D	T2D	Cirrosis	Cáncer Colorrectal	Obesidad	IBD
Vip	Nº comp.	1	1	2	1	1	2
	Punto de corte	0.3	0.3	0.4	0.4	0	0.35
	VIP	0.09	0.48	0.6	0.9	1.65	0.55
RC	Nº comp.	10	2	2	1	9	2
	Punto de corte	0.1	0.3	0.4	0.4	0.2	0.35
	Nº Var	120	20	60	80	20	100
SR	Nº comp.	1	2	2	1	2	2
	Punto de corte	0.3	0.35	0.3	0.4	0.3	0.35
	SR	0	0.0378	0.432	0.0104	0.004	0.0063
SPLS	Nº comp.	1	1	2	1	1	2
	Punto de corte	0.3	0.3	0.3	0.3	0.15	0.35
	eta	0	0.9	0.9	0.3	0.9	0.3
T2	Nº comp.	1	2	2	1	2	2
	Punto de corte	0.3	0.25	0.45	0.35	0	0.4
	alpha	1	0.2	0.7	0.25	0.05	0.8
BVE	Nº comp.	10	2	2	1	4	2
	Punto de corte	0.1	0.3	0.4	0.4	0.25	0.4
	VIP	0.72	0.9	0.54	0.84	0.98	0.56
PLS_DA	Nº comp.	1	1	2	1	2	2
	Punto de corte	0.3	0.3	0.45	0.35	0.35	0.35
RF	Punto de corte	0.5	0.35	0.4	0.4	0	0.35
Boruta	Punto de corte	0.35	0.35	0.35	0.4	0.1	0.25
VarSelRF	Punto de corte	0.4	0.3	0.35	0.35	0	0.3

Tabla 4.3: Modelo Lineal Mixto F1-Score

	SC	SCM	GL Var	GL Res	F-Ratio	P-valor	
Preprocesamiento	0.04	0.01	3	2237.14	30.53	2.4e-19	***
Técnica	0.35	0.04	9	2237.14	83.65	3.9e-134	***
BBDD	3.05	0.61	5	54.09	1324.41	3.1e-55	***
Preprocesamiento:Técnica	0.06	0.00	27	2237.14	4.72	1.6e-14	***
Preprocesamiento:BBDD	0.33	0.02	15	2237.14	48.34	1.6e-124	***
Técnica:BBDD	0.95	0.02	45	2237.14	46.00	1.6e-280	***

Nota: *** Significativo utilizando α= 0.05

modelos lineales mixtos, con los efectos fijos y aleatorios ya descritos, con el fin de determinar si existían diferencias estadísticamente significativas entre los modelos comparados.

4.2.1. F1-Score

Los resultados del modelo lineal mixto para el F1-score se pueden observar en la Tabla 4.3. Los efectos técnica, BBDD y preprocesamiento resultaron ser significativos, al igual que sus interacciones. Por lo tanto, estos factores influyen significativamente en la media del F1-score y los analizamos a continuación. Para comparar los niveles de los distintos factores, primero se calculan las MME y luego se aplica el test post-hoc de Tukey para realizar las comparaciones. Primero, compararemos el efecto de los modelos con y sin técnica de selección. La Figura 4.1 muestra que los valores del F1-score varían entre 0.71 y 0.76. Es decir, no existen diferencias considerables al variar el modelo con o sin técnica de selección de variables. Sin embargo, sí se observan diferencias estadísticamente significativas entre algunos de los métodos (intervalos determinados por las flechas rojas no solapantes), como comentaremos a continuación. El peor resultado se obtiene para el PLS-DA sin selección de variables. Cualquier otra técnica mejora significativamente al modelo PLS-DA con todas las variables. Por otra parte, el modelo RF sin selección de variables obtiene el mayor valor del F1-score, aunque sin diferencias significativas con RF-Boruta ($P = ,09$). Ambos son significativamente mejores que el resto de técnicas comparadas. En cuanto a los técnicas de selección de variables para PLS-DA, no se observan mayores diferencias significativas entre ellas, exceptuando la diferencia entre la peor (RC) y la mejor (BVE) ($P = ,003$). Dado el objetivo de nuestro estudio, es imprescindible poder seleccionar las bacterias asociadas a la enfermedad analizada por lo que deberemos sacrificar la capacidad predictiva del

modelo en aras de mejorar su interpretabilidad. Así, los modelos a considerar a partir de esta comparativa serían RF-Boruta (F1-score medio 0.752), BVE (0.736), SPLS (0.733) y VIP (0.733).

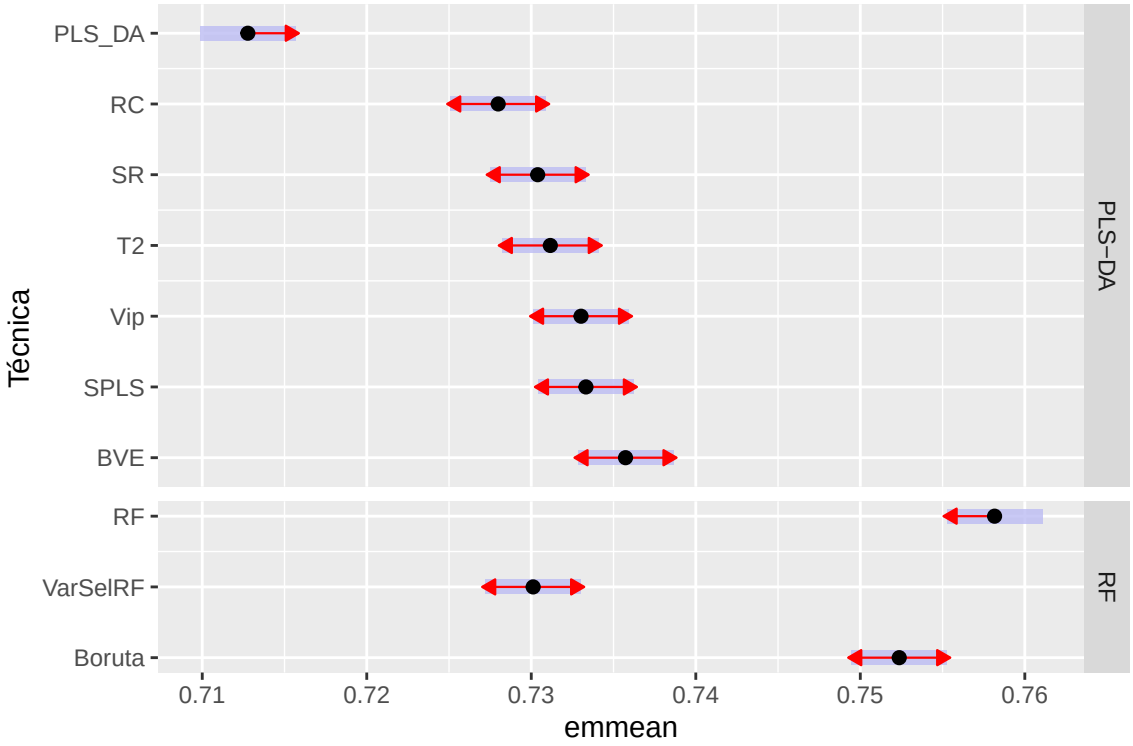

Figura 4.1: Comparación de las MME del F1-score respecto a la técnica de selección

El segundo efecto que se analiza es la BBDD. En la Figura 4.2 se puede observar que existen diferencias significativas entre todas las enfermedades estudiadas. El F1-score medio varía entre 0.662 y 0.825, siendo IBD y Cirrosis, la peor y mejor en F1-score medio respectivamente. Esto corrobora lo descrito en el estudio realizado anteriormente con estas enfermedades [29]. En contraste, la enfermedad Obesidad (F1-score 0.785) obtiene resultados considerablemente mejores respecto al estudio anterior y al de Pasolli [10].

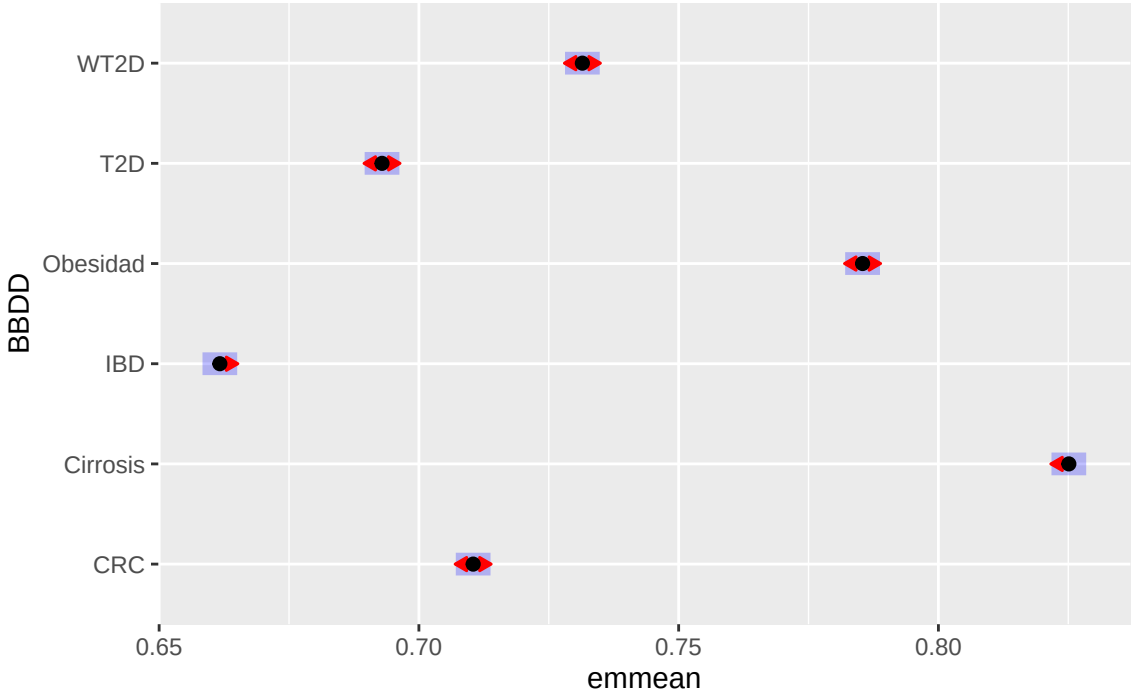

Figura 4.2: Comparación de las MME del F1-score respecto a la BBDD

Finalizando con el análisis de los efectos fijos simples, se analiza el efecto del preprocesamiento. En la Figura 4.3 se observa que el mejor preprocesamiento es el S2-N2 obteniendo un F1-score medio (0.741) y diferenciándose significativamente de los demás preprocesamientos. Entre los preprocesamientos S1-N1 y S1-N2 no existen diferencias estadísticamente significativas, por lo que aplicar la normalización N1 (TSS) o N2 (CLR) en combinación con la normalización S1 no implica diferencias significativas en los resultados. Por el contrario, entre los preprocesamientos S2-N1 y S2-N2 si existen diferencias significativas, por lo que con la normalización S2 la aplicación de N1 o N2 si afecta significativamente en la media del F1-score.

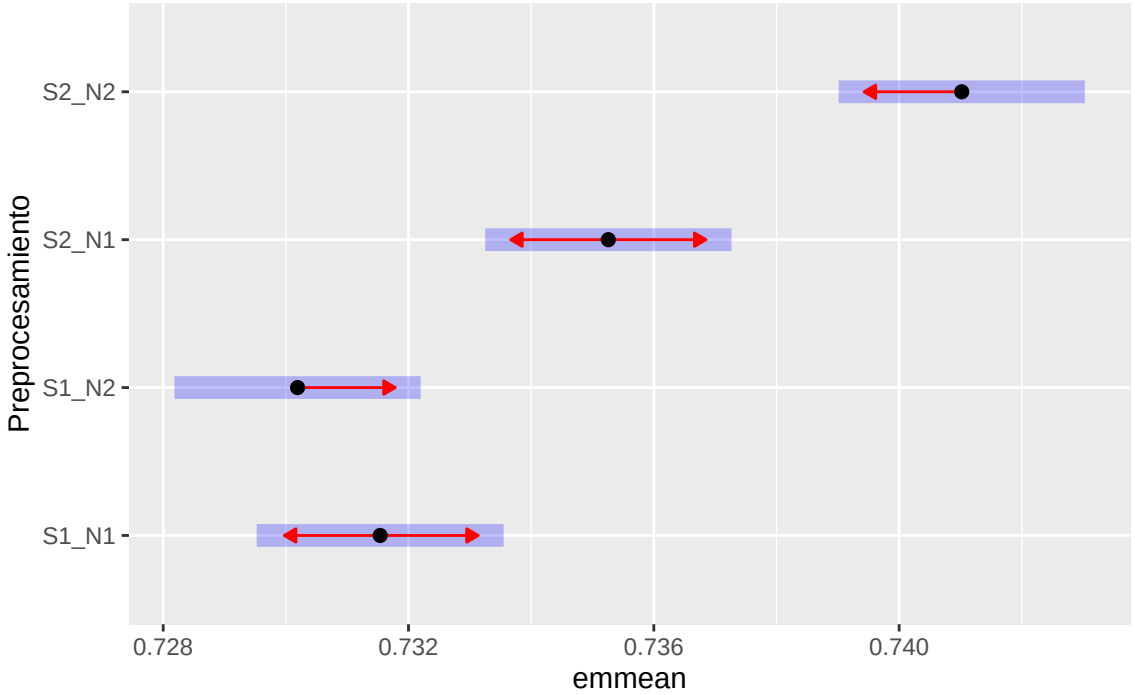

Figura 4.3: Comparación de las MME del F1-score respecto al preprocesamiento de los datos

A la vista de los resultados anteriores y teniendo en cuenta que las interacciones entre estos factores son también significativas, lo primero que cabe preguntarse es si para todas las BBDD analizadas y para todas las técnicas de preprocesamiento de los datos, se mantienen las mismas conclusiones sobre las mejores o peores técnicas de selección de variables. En la Figura 4.4 podemos observar la interacción entre la técnica y la BBDD. Aunque este gráfico es meramente descriptivo, el gráfico A.1 del Anexo A.4 nos ayudará a confirmar si las conclusiones que sacamos son o no estadísticamente significativas. En primer lugar, podemos ver que, en el caso de las BBDD de Cirrosis y CRC, RF-Boruta sigue siendo el mejor modelo con selección de variables (con diferencias estadísticamente significativas con respecto a los modelos PLS-DA con selección de variables). No obstante, apreciamos que no se puede afirmar lo mismo para el resto de BBDD, en las que los modelos PLS-DA con selección de variables presentan un F1-score medio similar o superior al de las técnicas RF (sin diferencias significativas). El único caso donde una técnica PLS-DA se diferencia significativamente de RF-Boruta es en la BBDD WT2D, donde RC tiene una media F1-score significativamente mejor que RF-Boruta.

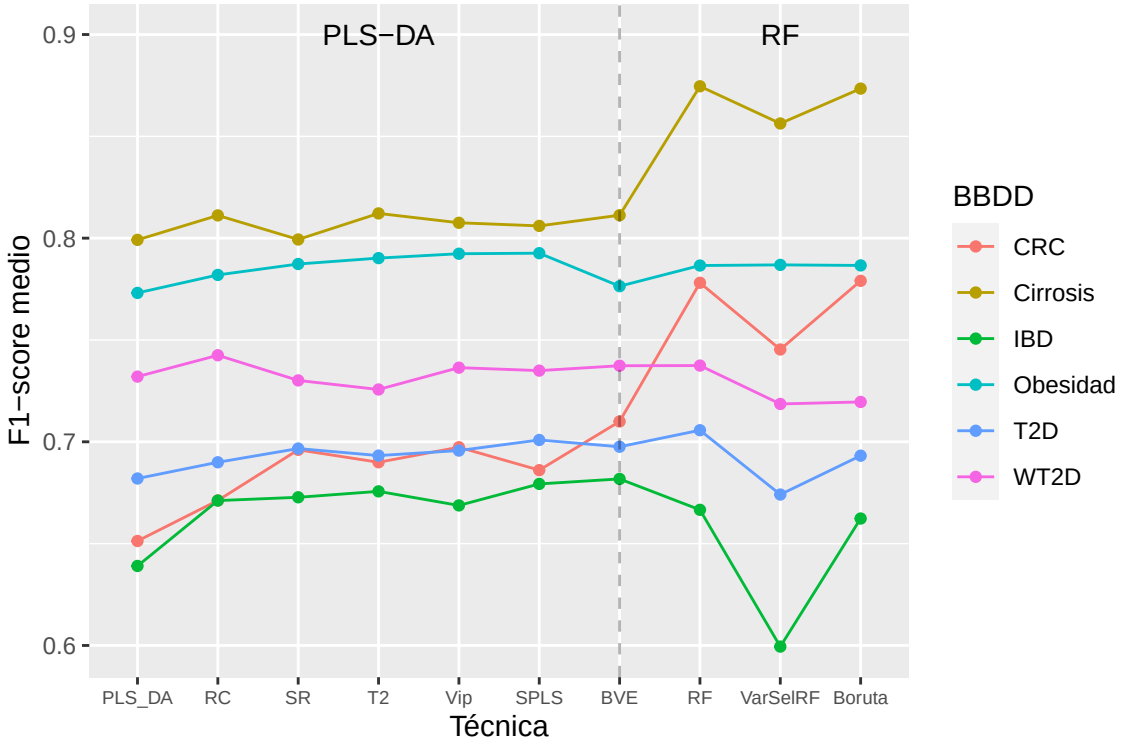

Figura 4.4: Interacción entre la técnica y BBDD

En cambio, a nivel de estrategia de preprocesamiento de los datos (Figura 4.5), parece mantenerse la superioridad de RF-Boruta respecto a las técnicas PLS-DA en cualquier estrategia de preprocesamiento. Entre los modelos PLS-DA, BVE sigue siendo la mejor técnica entre las de PLS-DA, excepto para el caso de S1-N2, en la que parece destacar SPLS. Con el fin de corroborar, se analiza el gráfico A.2 del Anexo A.4, donde se observa que RF-Boruta es significativamente mejor que las técnicas del PLS-DA, exceptuando a BVE que en los preprocesamientos S1-N1, S2-N1 y S2-N2 no tiene diferencias significativas con RF-Boruta. Además, particularmente en el preprocesamiento S2-N2 la técnica VIP tampoco posee diferencias significativas con RF-Boruta.

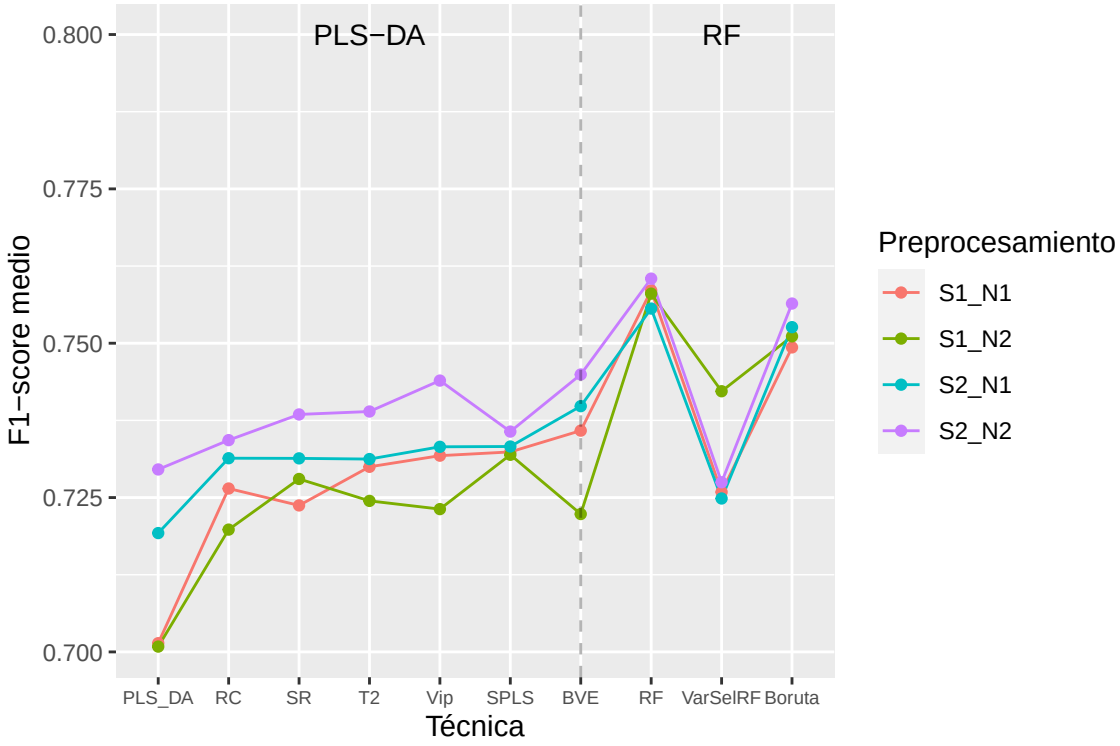

Figura 4.5: Interacción entre la técnica y el preprocesamiento de los datos

Por último, aunque a partir del estudio del efecto simple parecía claro que en general el mejor preprocesamiento era S2-N2, al estudiar la interacción con la BBDD (Figura 4.6), se puede apreciar que no existe un patrón claro del preprocesamiento óptimo a elegir, ya que dependerá de la BBDD que se esté analizando.

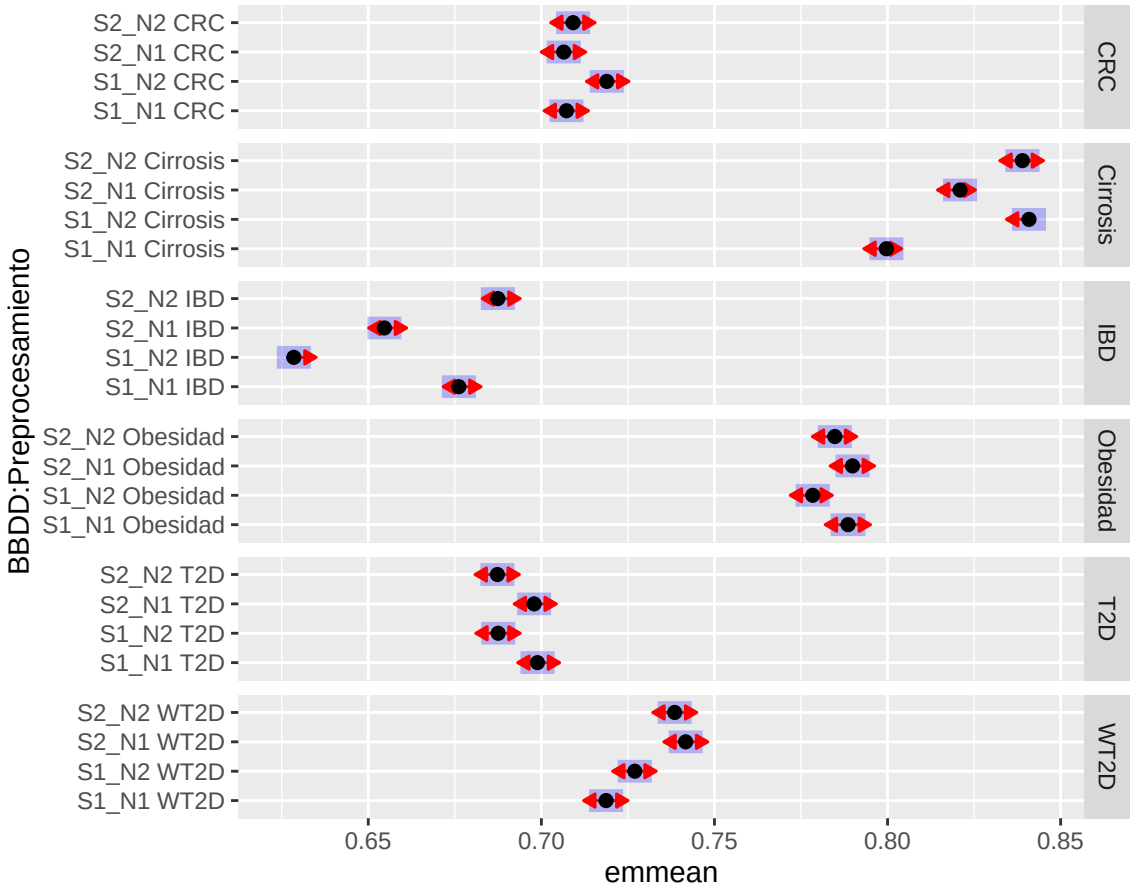

Figura 4.6: Comparación MME del F1-score de la interacción preprocesamiento y BBDD

Por tanto, concluimos que es complicado determinar una única estrategia de análisis que funcione bien en todos los casos en cuanto a la bondad de clasificación se refiere. No obstante, es importante también tener en cuenta otro aspecto al comparar técnicas de selección de variables: la estabilidad de dicha selección, que se estudia en la siguiente sección.

4.2.2. Estabilidad

En este caso, se ajustó un modelo de efectos fijos, ya que el efecto aleatorio presentaba baja variabilidad y producía problemas de singularidad en el ajuste. Además, como el objetivo era comparar la estabilidad de las variables seleccionadas, se descartaron los modelos PLS-DA y RF sin selección de variables.

Los resultados del modelo ajustado se pueden ver en la Tabla 4.4. Los efectos y sus interacciones son significativos por lo que se analizan en detalle a continuación. Las comparaciones de las MME según la técnica aplicada se observan en la Figura

Tabla 4.4: Modelo Lineal Estabilidad

Variable	GL Variable	GL Residuos	F-Ratio	P-valor	
Preprocesamiento	3	1832	433.83	6.6e-213	***
Técnica	7	1832	142.85	3.1e-168	***
BBDD	5	1832	92.59	4.5e-87	***
Preprocesamiento:Técnica	21	1832	19.76	8.3e-67	***
Preprocesamiento:BBDD	15	1832	24.06	2.4e-61	***
Técnica:BBDD	35	1832	39.19	3.9e-194	***

Nota: *** Significativo utilizando α= 0.05

4.7. El método RF-Boruta obtiene la mejor estabilidad media (0.912) seguido de SPLS (0.897), aunque no existen diferencias significativas entre ellos. Por el contrario el método VarSelRF es significativamente el peor de todos los métodos y tiene una estabilidad media de 0.784. Entre los métodos PLS-VIP, T^2, SR y BVE no existen diferencias significativas.

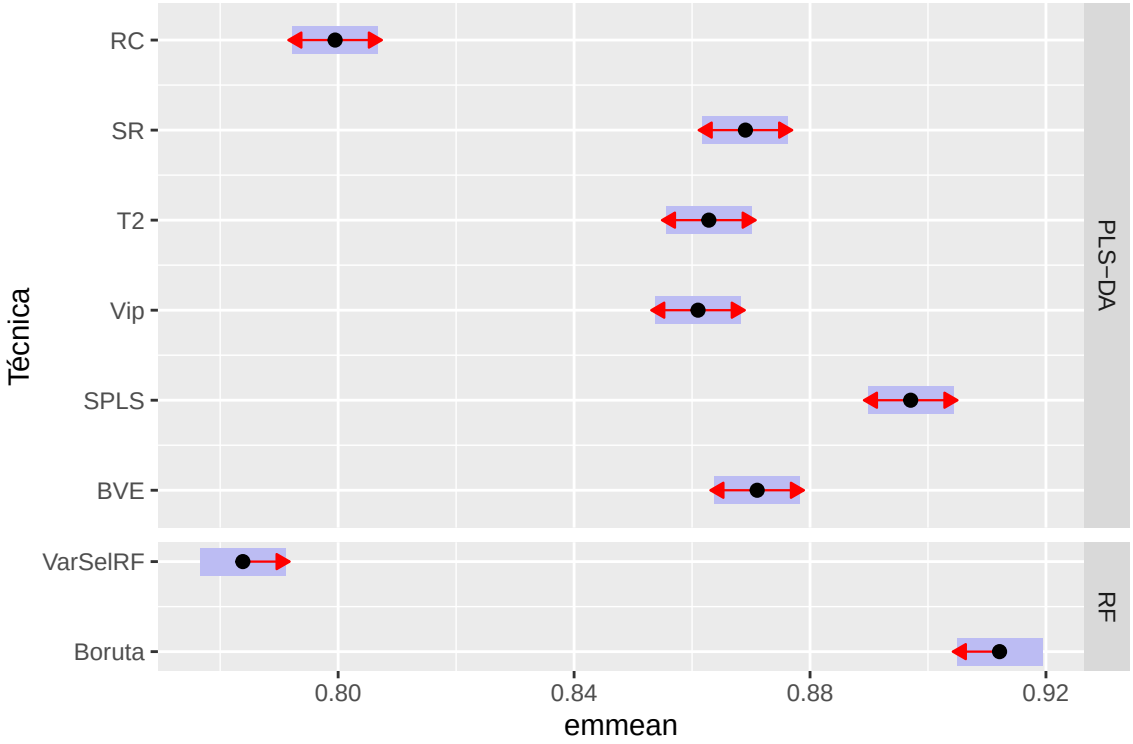

Figura 4.7: Comparación MME de la estabilidad respecto a la técnica de selección

En el caso del efecto simple de la BBDD (Figura 4.8), se observa que la estabilidad media varía entre 0.820 (IBD) y 0.892 (Obesidad). La enfermedad IBD es significativa-

mente peor que las demás. Por el contrario, Cirrosis y Obesidad son las enfermedades con mejor estabilidad media, aunque sin diferencias significativas entre ellas.

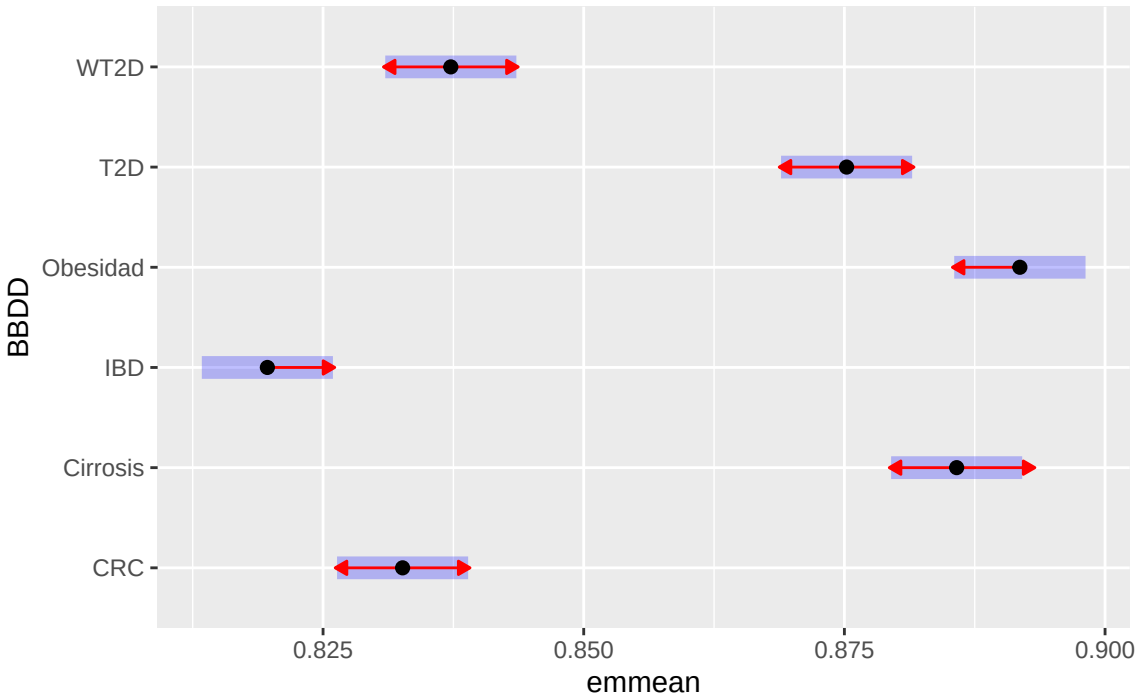

Figura 4.8: Comparación MME de la estabilidad respecto a la BBDD

El último efecto simple, el efecto de preprocesamiento (Figura 4.9), se observa que el preprocesamiento S1-N1 se diferencia significativamente del resto con una estabilidad media de 0.943. Es decir, no aplicar el filtro de prevalencia (S1) y la normalización TSS (N1) contribuye a la estabilidad de los métodos aplicados. Por el contrario, al aplicar la normalización CLR (N2) la estabilidad media disminuye significativamente.

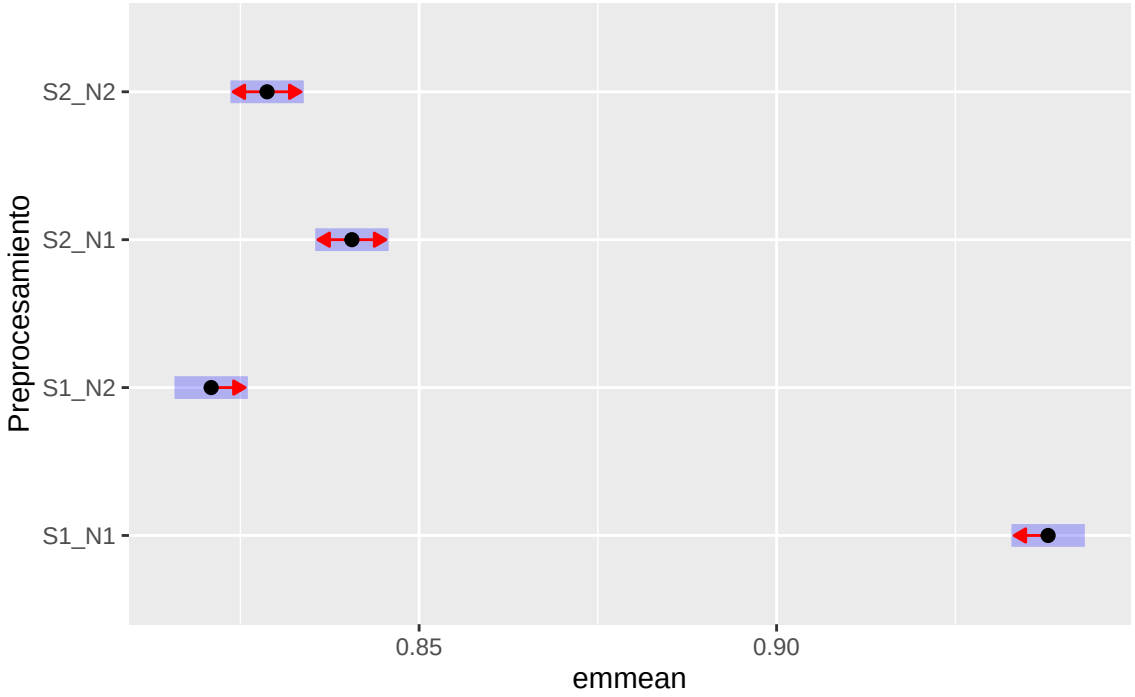

Figura 4.9: Comparación MME de la estabilidad respecto al preprocesamiento

A continuación, se analizan las interacciones ya que también son significativas. En el caso de la interacción entre técnica y BBDD (Figura 4.10) se observa que no hay una técnica de selección de variables que domine en todas las BBDD. Sin embargo, pareciera que la técnica VarSelRF en 4 de las 6 BBDD es la peor de las técnicas de selección de variables. Para corroborar si estas diferencias son significativas, se utilizó el gráfico A.3 del Anexo A.4, donde se observa que la técnica VarSelRF es la peor significativamente solo en la BBDD T2D.

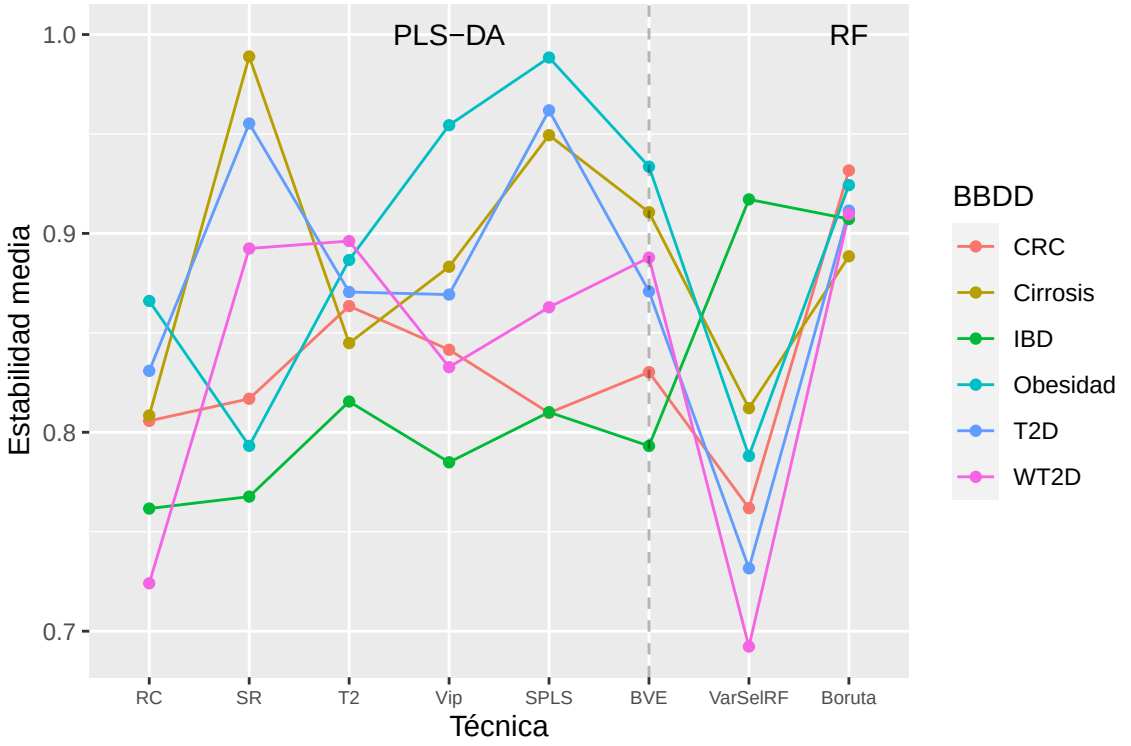

Figura 4.10: Interacción entre técnica y BBDD (Estabilidad)

Para la interacción entre la técnica y el preprocesamiento aplicado a los datos (Figura 4.11), se observa que todas las técnicas de selección de variables obtienen las mejor estabilidad media con el preprocesamiento S1-N1. Al analizar el gráfico A.4 del Anexo A.4 se corrobora que esta diferencia es significativa, exceptuando para la técnica RF-Boruta donde el preprocesamiento S1-N2 no se diferencia significativamente del preprocesamiento S1-N1. Otra técnica que destaca es SPLS en el preprocesamiento S2-N2 (estabilidad media = 0.910), alcanzando las estabilidades medias del preprocesamiento S1-N1.

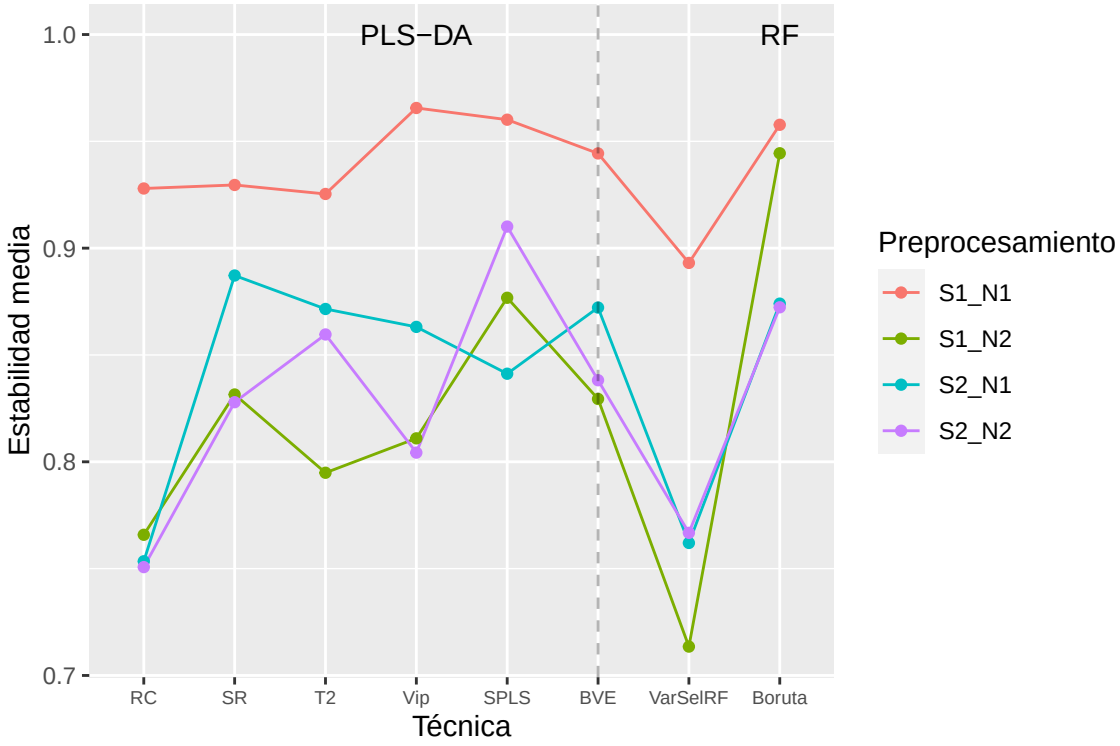

Figura 4.11: Interacción entre técnica y preprocesamiento de los datos (Estabilidad)

Finalmente, se analiza la interacción entre la BBDD y el preprocesamiento (Figura 4.12), donde se observa que para las bases de datos de CRC, Cirrosis, IBD, Obesidad y T2D el preprocesamiento S1-N1 es significativamente mejor que el resto de los preprocesamientos. En el caso de WT2D, los preprocesamientos S1-N1, S2-N1 y S2-N2 presentan estabilidades medias similares, y no se observan diferencias significativas entre ellos.

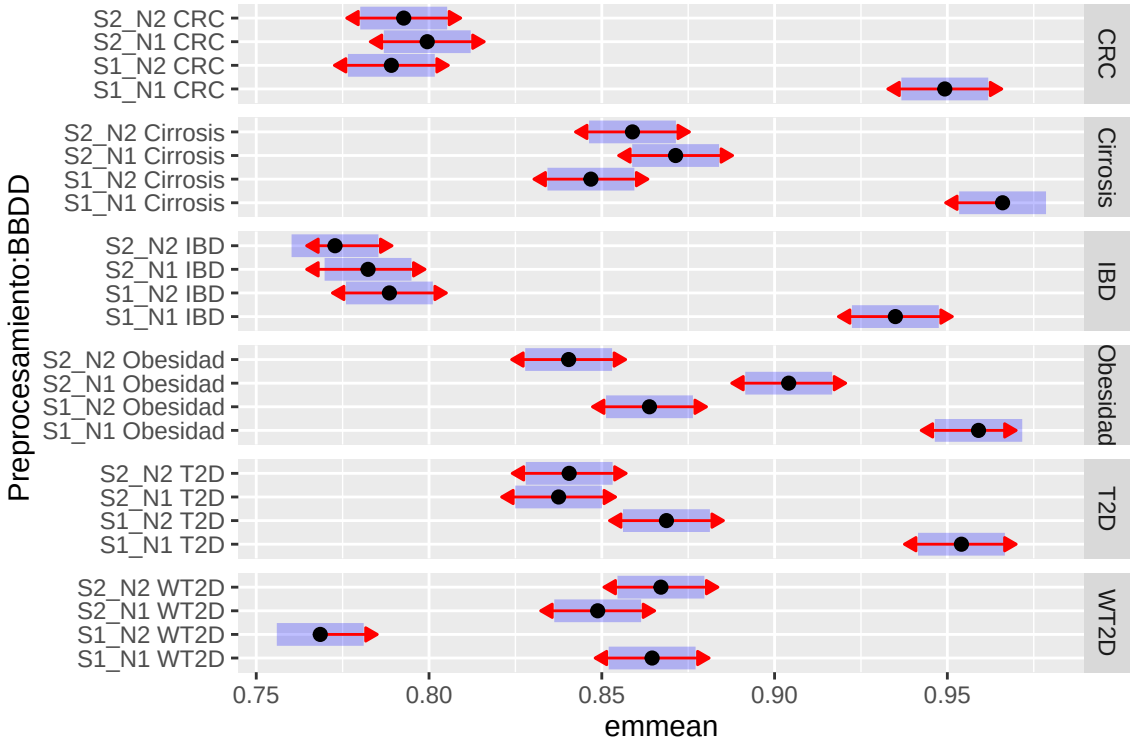

Figura 4.12: Interacción entre BBDD y preprocesamiento de los datos (Estabilidad)

4.3. Interpretación de los modelos seleccionados

Considerando que se han ajustado 240 modelos en total, elegiremos solamente algunos de ellos para ilustrar la interpretación de los resultados. En este caso, hemos escogido la BBDD de Obesidad, ya que es una de las que presenta mayor estabilidad en la selección de variables y nos permitirá una mejor interpretación de las variables seleccionadas. Se analizarán las técnicas de selección de variables y el preprocesamiento que mejores resultados obtuvieron tanto para el F1-score medio como la estabilidad media. Además, se selecciona al menos una técnica de selección proveniente del PLS-DA y otra del RF para comparar distintos modelos base. Por esto, se compararán las técnicas de selección de variables RF-Boruta, SPLS y VIP con el preprocesamiento S2-N1, ya que son los métodos y uno de los preprocesamientos que mejor equilibrio presentan entre F1-score y estabilidad (Figura 4.13).

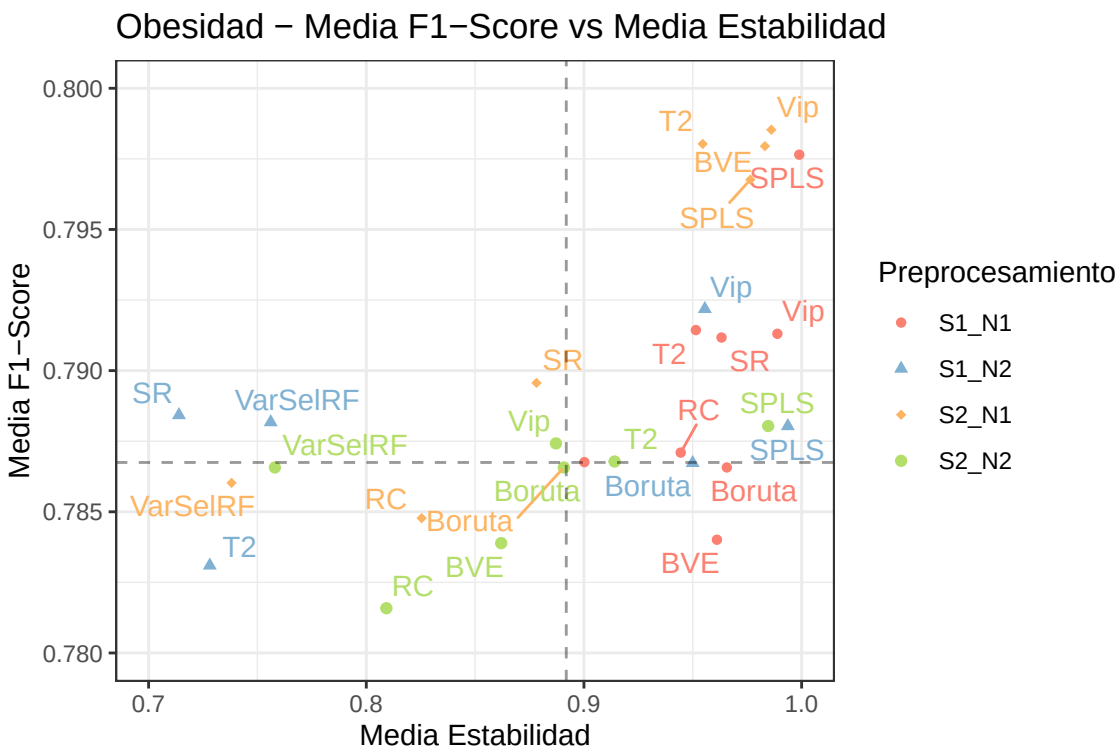

Figura 4.13: Obesidad - media F1-Score vs estabilidad media

Número de variables seleccionadas

Al realizar el ajuste mediante validación cruzada repetida, el número de variables seleccionadas varía entre los distintos *folds*. Para la enfermedad Obesidad con el preprocesamiento S2-N1 y los métodos seleccionados esto se puede observar en la Figura 4.14. Al ordenar los métodos por la mediana del número de variables seleccionadas, observamos que la técnica RF-Boruta es la que más selecciona con una mediana de 10 variables. También, observamos que las técnicas de selección VIP y SPLS realizan un filtrado muy exigente con una mediana de 4 variables para ambas técnicas. Un dato importante a considerar es que el número de variables sin selección de esta BBDD es de 155 predictores, por lo que los modelos con selección de variables utilizan menos del 7 % de los predictores originales, sin que existan diferencias significativas en la bondad de clasificación respecto a los modelos sin selección (Figura A.5 del Anexo A.5). El mínimo y máximo número de variables seleccionadas por estos métodos se pueden ver en la Tabla A.5 del Anexo A.5.

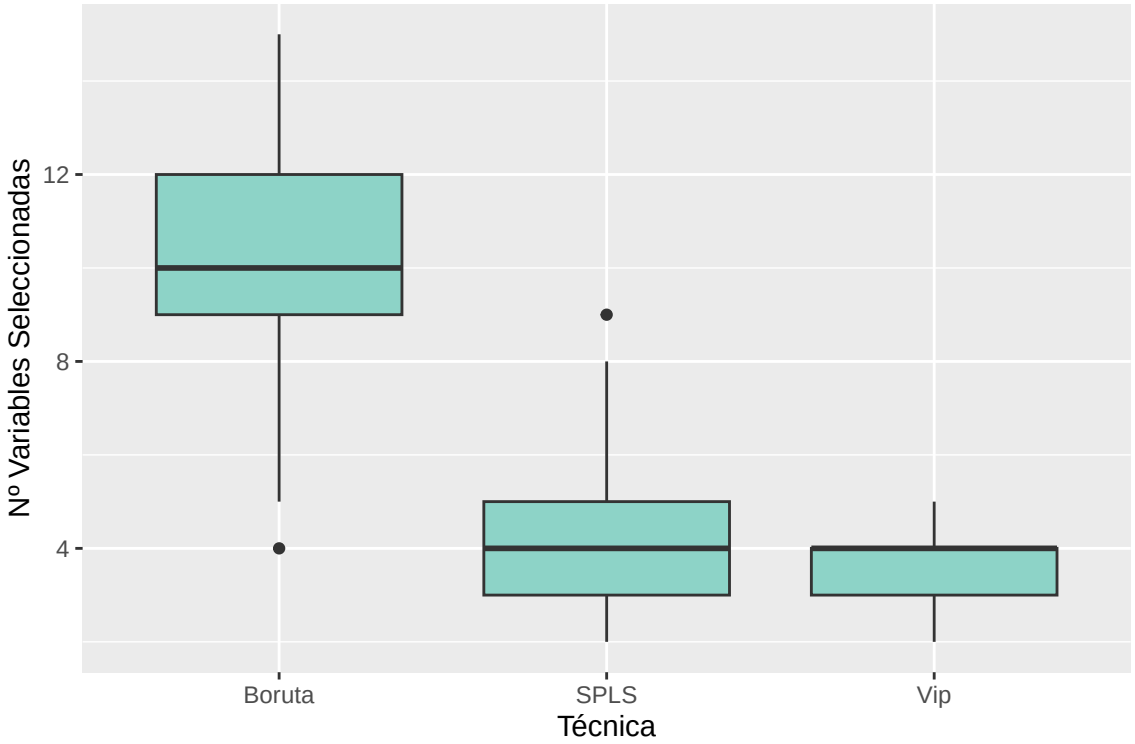

Figura 4.14: Nº Variables seleccionadas por técnica en cada fold y repetición de la CV

Variables seleccionadas

Con el fin de obtener un conjunto fijo de variables y poder comparar las especies de bacterias seleccionadas por los distintos modelos, se seleccionan las especies que al menos hayan sido seleccionadas un 50 % de las veces. En este caso, al tener 100 folds en total (10 repeticiones con 10 folds) se seleccionan las especies que hayan sido seleccionadas al menos 50 veces. Un ejemplo se puede ver en la Figura 4.15 donde para el método RF-Boruta 7 especies de bacterias fueron seleccionadas al menos 50 veces.

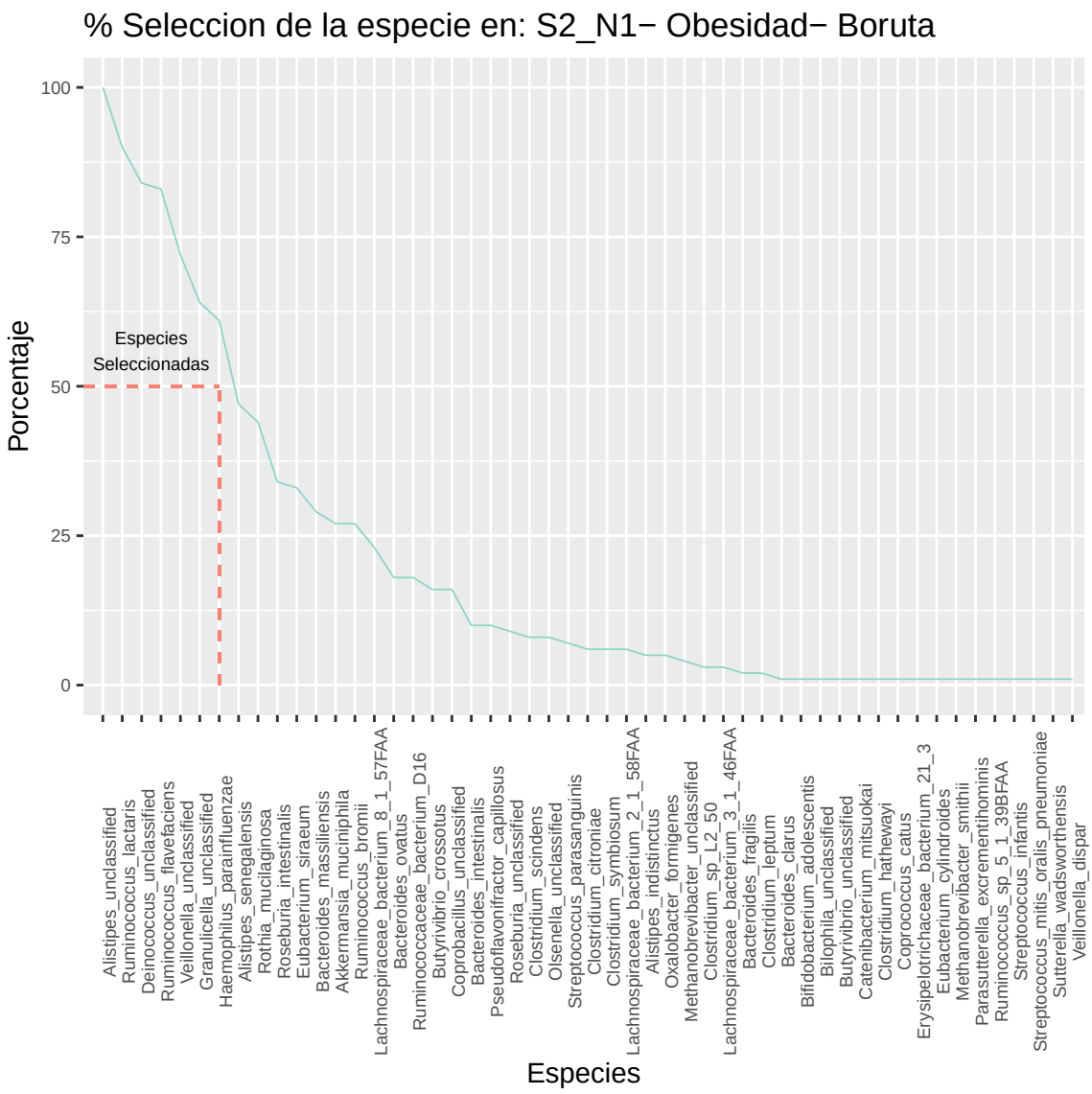

Figura 4.15: Ejemplo de especies seleccionadas según su porcentaje de selección para la BBDD Obesidad con preprocesamiento S2-N1 según el método RF-Boruta

Aplicando lo mismo para las técnicas RF-Boruta, SPLS y VIP se obtienen las especies de bacterias seleccionadas para cada uno de los métodos (Tabla 4.5). Al comparar las especies seleccionadas por los tres métodos (Figura 4.16) se observa que los tres métodos seleccionan las especies *Alistipes unclassified, Deinococcus unclassified, Ruminococcus flavefaciens*. Además, entre SPLS y VIP no existen diferencias en las especies seleccionadas. Sin embargo, RF-Boruta se diferencia de SPLS y VIP al no seleccionar la especie *Eubacterium_siraeum*, pero sí selecciona otras que los métodos SPLS y VIP no seleccionaron (*Ruminococcus lactaris, Veillonella unclassified, Granulicella unclassified, Haemophilus parainfluenzae*).

Tabla 4.5: Especies seleccionadas por técnica

Técnica	Nº	Especies
Boruta	7	Alistipes_unclassified, Deinococcus_unclassified, Granulicella_unclassified, Haemophilus_parainfluenzae, Ruminococcus_flavefaciens, Ruminococcus_lactaris, Veillonella_unclassified.
SPLS	4	Alistipes_unclassified, Deinococcus_unclassified, Eubacterium_siraeum, Ruminococcus_flavefaciens.
VIP	4	Alistipes_unclassified, Deinococcus_unclassified, Eubacterium_siraeum, Ruminococcus_flavefaciens.

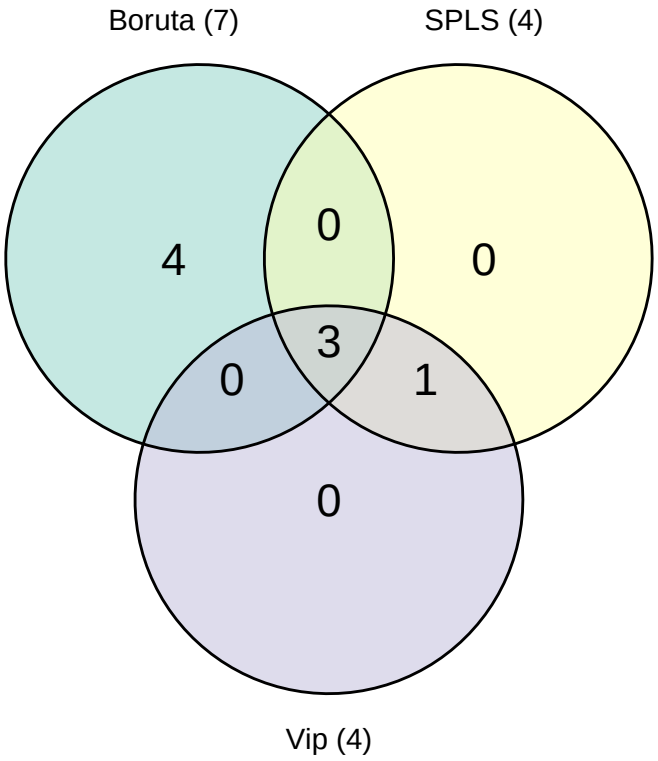

Figura 4.16: Comparación de especies de bacterias seleccionadas por los métodos RF-Boruta, SPLS y VIP

Al revisar en la literatura las especies de bacterias seleccionadas por los tres métodos encontramos lo descrito a continuación. Se asocia la baja abundancia de la especie *Alistipes unclassified* en los individuos con obesidad respecto a los que no [70] [71] [72] [73]. Para la especie *Ruminococcus flavefaciens*, un estudio ha encontrado una baja abundancia de esta especie en los casos obesos [74]. Otro estudio relacionado al síndrome metabólico (SM), condición muy ligada a la obesidad también confirma esta tendencia, encontrando una abundancia relativa menor en los casos con SM respecto a los casos de control [75]. Un estudio en animales (ratas) sobre el efecto del ejercicio en la microbiota, reveló también que la especie *Ruminococcus flavefaciens* tenía una

abundancia relativa menor en las ratas obesas en comparación a los otros dos grupos no obesos [76]. Para el caso de la especie *Deinococcus* no se encontraron estudios que relacionen esta especie con la enfermedad, por lo que esta especie es candidata a ser analizada en futuros estudios para confirmar si es una hallazgo relevante de nuestro análisis.

En el caso de la especie *Eubacterium_siraeum*, la cual fue seleccionada por VIP y SPLS, pero no por RF-Boruta, se asocia una mayor abundancia en los individuos no obesos [77] [78] [79].

Respecto a las especies de bacterias que solo fueron seleccionadas por el modelo RF-Boruta, específicamente *Ruminococcus_lactaris*, *Veillonella_unclassified*, *Granulicella_unclassified* y *Haemophilus_parainfluenzae* no se encontraron estudios que evidencien alguna asociación directa con la enfermedad Obesidad.

Por otra parte, una de las especies que en la literatura aparece frecuentemente asociada a los estudios (en humanos y animales) sobre la obesidad es la especie *Akkermansia muciniphila*, ya que su abundancia se correlaciona negativamente con la obesidad [80] [81] y también se ha asociado su aumento en abundancia en dietas de pérdida de peso [82], sin embargo en ninguno de los métodos analizados se seleccionó esta especie.

Para observar las relaciones entre las especies seleccionadas y la enfermedad, y ver si coinciden con lo comentado anteriormente, se realiza el gráfico de weights w^*c (Figura 4.17) con la técnica de selección VIP (que es idéntico al que se obtendría con SPLS puesto que ambos seleccionan las mismas variables). Se puede observar claramente que las especies seleccionadas se posicionan en el extremo opuesto a la presencia de la enfermedad, por lo que existe una correlación negativa entre la abundancia de las especies (*Alistipes unclassified*, *Ruminococcus flavefaciens* y *Eubacterium_siraeum*) y la obesidad, en consonancia con lo descrito en la literatura. Además podemos observar una estrecha correlación entre las especies *Alistipes* y *Deinococcus*, que también resultaría interesante estudiar. Se puede apreciar en este gráfico y en el de scores (Figura 4.18) la ventaja interpretativa del modelo PLS frente a RF, ya que este último nos permite seleccionar las variables más relevantes pero no su relación entre ellas y con la enfermedad. Se requerirían otros análisis adicionales para indagar más sobre dichas relaciones.

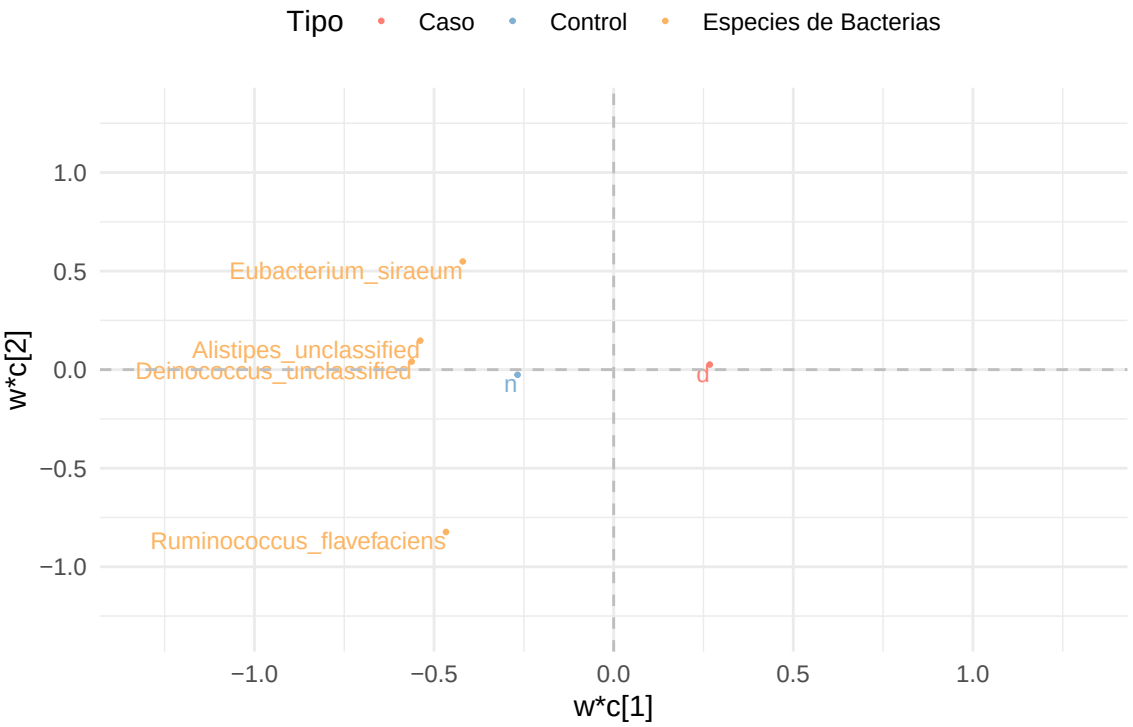

Figura 4.17: Gráfico de w*c Obesidad - VIP y SPLS

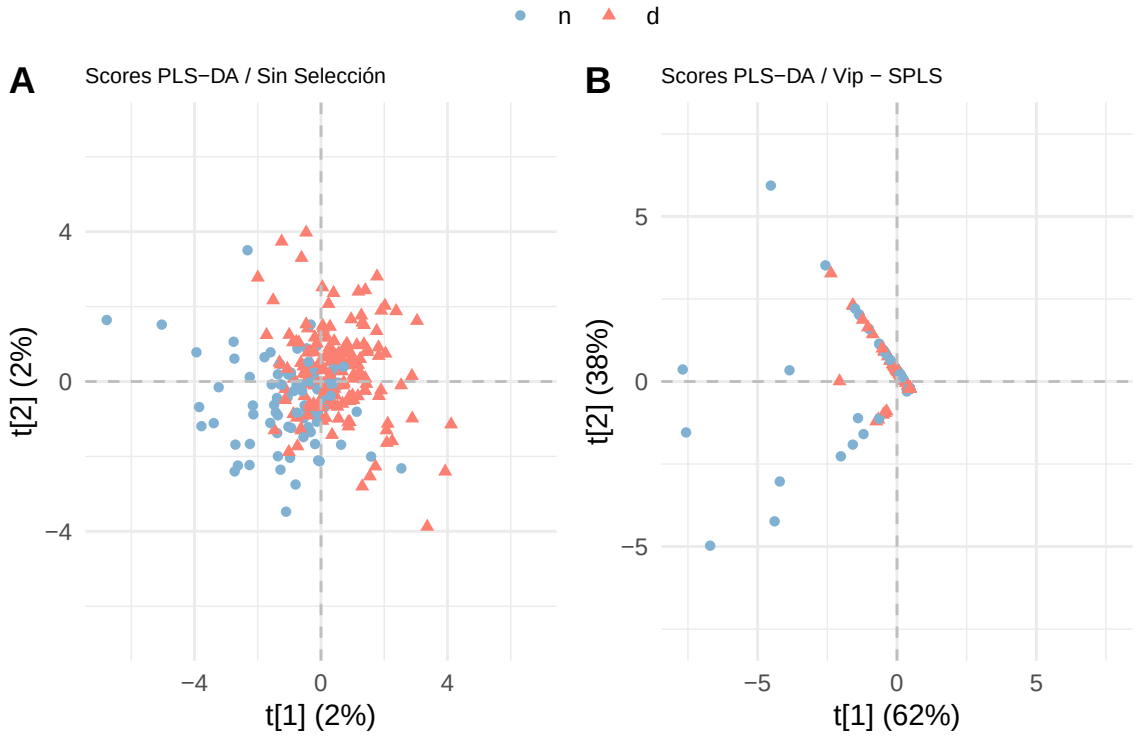

Figura 4.18: Gráfico de Scores del PLS-DA antes y después de las selección de variables

Para concluir el análisis de esta BBDD, hay que recordar que en los resultados del MLM del F1-score se observó que la interacción técnica y BBDD fue significativa, y particularmente en la BBDD de Obesidad no existían diferencias significativas entre los modelos con selección y sin selección, en consecuencia las técnicas de selección de variables en esta BBDD no mejoran significativamente la capacidad predictiva de los modelos sin selección, pero sí nos ayudan a reducir la complejidad del modelo y mejorar su interpretación, especialmente en las técnicas de selección del PLS-DA.

Capítulo 5

Conclusiones

El foco principal de este trabajo fue la aplicación y comparación de técnicas de selección de variables en el modelo PLS-DA aplicado a datos de microbioma. La hipótesis a validar fue que la aplicación de estas técnicas de selección de variables mejoraría el rendimiento e interpretación de los modelos ajustados. Para validar esto, se comenzó realizando una revisión bibliográfica de las técnicas de selección de variables del PLS, seleccionando aquellas que mejores resultados obtuvieron en los estudios revisados para ser implementadas en el PLS-DA. Luego, estas técnicas de selección se aplicaron sobre conjuntos de datos de microbioma de distintas patologías (CRC, Cirrosis, IBD, Obesidad y Diabetes tipo 2), que fueron recopilados en un estudio anterior [10] y a los cuales se les aplicaron distintas estrategias de preprocesamiento [29]. Además, se compararon estas técnicas del PLS-DA con el modelo RF y técnicas de selección del RF, para así obtener un mejor marco de comparación y contrastarlo con uno de los modelos más utilizados en este contexto. En total se ajustaron 240 modelos y al compararlos las conclusiones son las siguientes:

- En las BBDD estudiadas, aplicar técnicas de selección de variables del PLS-DA no supone una disminución significativa del F1-score medio respecto al PLS-DA sin selección, incluso en algunas BBDD (CRC e IBD) implica una mejora significativa del F1-score medio. Es decir, en el peor de los casos se obtiene interpretabilidad sin perder capacidad predictiva de manera significativa. Por otra parte, el modelo RF y sus técnicas de selección (RF-Boruta y VarSelRF) tienen un F1-score medio significativamente mejor que las técnicas del PLS-DA solo en las BBDD CRC y Cirrosis. Respecto al preprocesamiento aplicado a los datos, no se puede proporcionar una recomendación, ya que depende de la técnica y la BBDD que se este analizando.

- Otro aspecto estudiado al comparar los modelos fue la estabilidad, es decir, cómo de consistentes son los modelos seleccionando las variables aunque cambien los datos de entrenamiento. La elección de la técnica de selección que maximice la estabilidad media dependerá de la BBDD analizada. Además, se recomienda aplicar el preprocesamiento S1-N1, ya que en todas las BBDD y técnicas de selección se obtendrán la estabilidades medias más altas.

- Al aplicar las técnicas de selección del PLS-DA a otras BBDD, será fundamental una estructura de validación adecuada (CV u otra) que permita comparar los modelos (con su respectiva técnica de selección y estrategia de preprocesamiento) para la elección del modelo óptimo para esa BBDD. También, esta estructura de validación será un punto central para la elección de los hiperparámetros óptimos que deben ser optimizados para cada modelo y técnica de selección.

Al realizar la interpretación de la BBDD de Obesidad con las mejores técnicas de selección se realizaron los siguientes hallazgos:

- Las técnicas de selección de variables reducen en gran medida la complejidad del modelo, utilizando muy pocas especies de bacterias como predictores de la enfermedad sin perder bondad de clasificación de manera significativa.

- La estabilidad es un aspecto fundamental a evaluar en una técnica de selección de variables. Técnicas más estables permitirán encontrar a los predictores más relevantes y generar una interpretación adecuada que no dependa en gran medida del conjunto de entrenamiento.

- Existe evidencia en la literatura de que las especies seleccionadas por las técnicas de selección aplicadas se asocian a la enfermedad estudiada. Además, las técnicas del PLS-DA nos permiten ver la relación entre las especies y la enfermedad aportando un potencial interpretativo respecto al RF y sus técnicas de selección.

En este estudio, existen varios puntos de mejora o alternativas que podrían ser implementados por otros investigadores. Por ejemplo, la elección de los hiperparámetros óptimos se hizo maximizando el F1-score, pero la elección se podría realizar optimizando otro tipo de métrica (AUC, Coeficiente de correlación de Matthews, Estabilidad, entre otros). Además, al variar la selección de variables por cada conjunto de entrenamiento utilizado, se decidió en este caso utilizar las variables que fueron seleccionadas al menos un 50 % del total de veces, pero este límite podría ser más o menos exigente. Finalmente, un punto importante y que podría ser medido en futuros trabajos tiene

que ver con el costo computacional de las técnicas de selección, ya que dependiendo del tipo de técnica (*"filter"*, *"wrapper"* y *"embedded"*) y la estructura de la validación cruzada los tiempos de ejecución pueden variar considerablemente.

Finalizando, podemos decir que el trabajo cumple los objetivos planteados de manera satisfactoria, ofreciendo una alternativa viable a los modelos comúnmente utilizados en el análisis de datos de microbioma, como es el caso del modelo Random Forest que, aunque en algunos casos pueda presentar una capacidad predictiva superior, está lejos de proporcionar la interpretabilidad de los modelos PLS.

Apéndice A

Anexos

A.1. Relación del Trabajo con los Objetivos de Desarrollo Sostenible de la Agenda 2030

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.	X			
ODS 4. Educación de calidad.				X
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico.				X
ODS 9. Industria, innovación e infraestructuras.			X	
ODS 10. Reducción de las desigualdades.				X
ODS 11. Ciudades y comunidades sostenibles.				X
ODS 12. Producción y consumo responsables.				X
ODS 13. Acción por el clima.				X
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.				X

Este trabajo se relaciona con el objetivo de desarrollo sostenible “Salud y Bienestar”, ya que compara el desempeño de modelos estadísticos, con y sin selección de variables, aplicados al análisis de datos de microbioma, con el fin de obtener modelos interpretables que permitan identificar biomarcadores potenciales para la prevención, el diagnóstico y el tratamiento de enfermedades. Además, se vincula con el objetivo “Industria, Innovación e Infraestructuras” al aplicar y comparar técnicas de selección

de variables dentro de un modelo específico (PLS-DA), aportando un enfoque distinto a los habitualmente utilizados, que otros investigadores podrían emplear en este tipo de análisis.

A.2. Parámetros Óptimos

Tabla A.1: Parámetros óptimos para normalización S1-N1

Técnica	Parámetro	WT2D	T2D	Cirrosis	Cáncer Colorrectal	Obesidad	IBD
Vip	Nº comp.	1	1	3	1	1	1
	Punto de corte	0.05	0.35	0.4	0.3	0	0.3
	VIP	1.05	1.87	1.8	1.9	2.04	0.88
RC	Nº comp.	1	2	2	2	1	1
	Punto de corte	0	0.3	0.35	0.25	0	0.3
	Nº Var	60	40	40	20	20	80
SR	Nº comp.	6	2	2	1	1	1
	Punto de corte	0.05	0.3	0.4	0.35	0.25	0.3
	SR	0.0053	0.062	0	0.021	0.0315	0
SPLS	Nº comp.	1	4	2	3	1	1
	Punto de corte	0.05	0.35	0.45	0.3	0.4	0.3
	eta	0.5	0.9	0.8	0.9	0.9	0.2
T2	Nº comp.	2	3	5	2	5	1
	Punto de corte	0	0.3	0.4	0.4	0.1	0.3
	alpha	0.4	0.25	0.05	0.2	0.05	0.45
BVE	Nº comp.	3	1	2	4	2	1
	Punto de corte	0	0.35	0.4	0.3	0.1	0.3
	VIP	0.78	0.84	2.16	0.91	0.98	0.64
PLS_DA	Nº comp.	1	1	2	1	1	1
	Punto de corte	0.3	0.3	0.4	0.35	0	0.3
RF	Punto de corte	0.5	0.35	0.4	0.4	0	0.35
Boruta	Punto de corte	0.35	0.4	0.4	0.45	0.1	0.35
VarSelRF	Punto de corte	0.05	0.3	0.45	0.4	0.05	0.45

Tabla A.2: Parámetros óptimos para normalización S1-N2

Técnica	Parámetro	WT2D	T2D	Cirrosis	Cáncer Colorrectal	Obesidad	IBD
Vip	Nº comp.	1	6	3	1	5	4
	Punto de corte	0.35	0.25	0.5	0.3	0.45	0.3
	VIP	0.45	1.4	0.8	0.9	2.04	0.6
RC	Nº comp.	1	1	3	1	4	2
	Punto de corte	0.4	0.35	0.4	0.3	0.35	0.25
	Nº Var	240	460	140	320	40	200
SR	Nº comp.	1	4	3	1	3	3
	Punto de corte	0.4	0.35	0.3	0.35	0.35	0.3
	SR	0.0344	0.0488	0.3888	0.0162	0.0074	0.0116
SPLS	Nº comp.	1	1	2	1	1	1
	Punto de corte	0.35	0.3	0.4	0.3	0.15	0.25
	eta	0.1	0.9	0.5	0.3	0.9	0.3
T2	Nº comp.	1	5	3	1	4	1
	Punto de corte	0.35	0.3	0.45	0.4	0.15	0.25
	alpha	0.65	0.05	0.5	0.2	0.3	0.45
BVE	Nº comp.	1	1	3	1	4	4
	Punto de corte	0.35	0.35	0.5	0.3	0.45	0.3
	VIP	0.42	0.88	0.66	0.77	2.07	0.56
PLS_DA	Nº comp.	1	2	1	1	2	1
	Punto de corte	0.35	0.3	0.35	0.35	0	0.25
RF	Punto de corte	0.5	0.4	0.5	0.4	0	0.35
Boruta	Punto de corte	0.4	0.35	0.35	0.45	0.15	0.35
VarSelRF	Punto de corte	0.4	0.35	0.4	0.4	0.15	0.4

Tabla A.3: Parámetros óptimos para normalización S2-N1

Técnica	Parámetro	WT2D	T2D	Cirrosis	Cáncer Colorrectal	Obesidad	IBD
Vip	Nº comp.	1	1	2	2	1	2
	Punto de corte	0.35	0.35	0.4	0.3	0.25	0.35
	VIP	0.18	1.5	0.27	1.08	2.31	0.6
RC	Nº comp.	1	1	2	2	4	2
	Punto de corte	0.35	0.3	0.35	0.35	0.05	0.3
	Nº Var	120	80	140	40	20	120
SR	Nº comp.	1	3	2	1	1	1
	Punto de corte	0.35	0.3	0.4	0.35	0.3	0.25
	SR	0	0.0326	0	0.02	0.0345	0.0084
SPLS	Nº comp.	1	5	2	1	1	2
	Punto de corte	0.35	0.35	0.4	0.3	0.25	0.3
	eta	0.1	0.9	0.2	0.3	0.7	0.3
T2	Nº comp.	3	2	2	2	1	2
	Punto de corte	0.2	0.3	0.35	0.3	0.25	0.35
	alpha	1	0.25	0.7	0.2	0.05	0.85
BVE	Nº comp.	1	1	2	3	1	2
	Punto de corte	0.35	0.3	0.35	0.3	0.25	0.35
	VIP	0.24	0.7	1.08	0.72	0.98	0.56
PLS_DA	Nº comp.	1	1	2	1	2	2
	Punto de corte	0.35	0.3	0.4	0.25	0	0.35
RF	Punto de corte	0.45	0.4	0.55	0.4	0	0.35
Boruta	Punto de corte	0.35	0.3	0.35	0.45	0	0.35
VarSelRF	Punto de corte	0.4	0.3	0.45	0.4	0.15	0.3

A.3. Detalle Casos sin Ajuste

Tabla A.4: Casos que no se ajustaron

norm	method	bbdd	f1	rep
S1_N1	T2	Obesidad	NA	1
S1_N1	T2	Obesidad	NA	2
S1_N1	T2	Obesidad	NA	3
S1_N1	T2	Obesidad	NA	6

A.4. Gráficos de comparaciones por pares F1-score

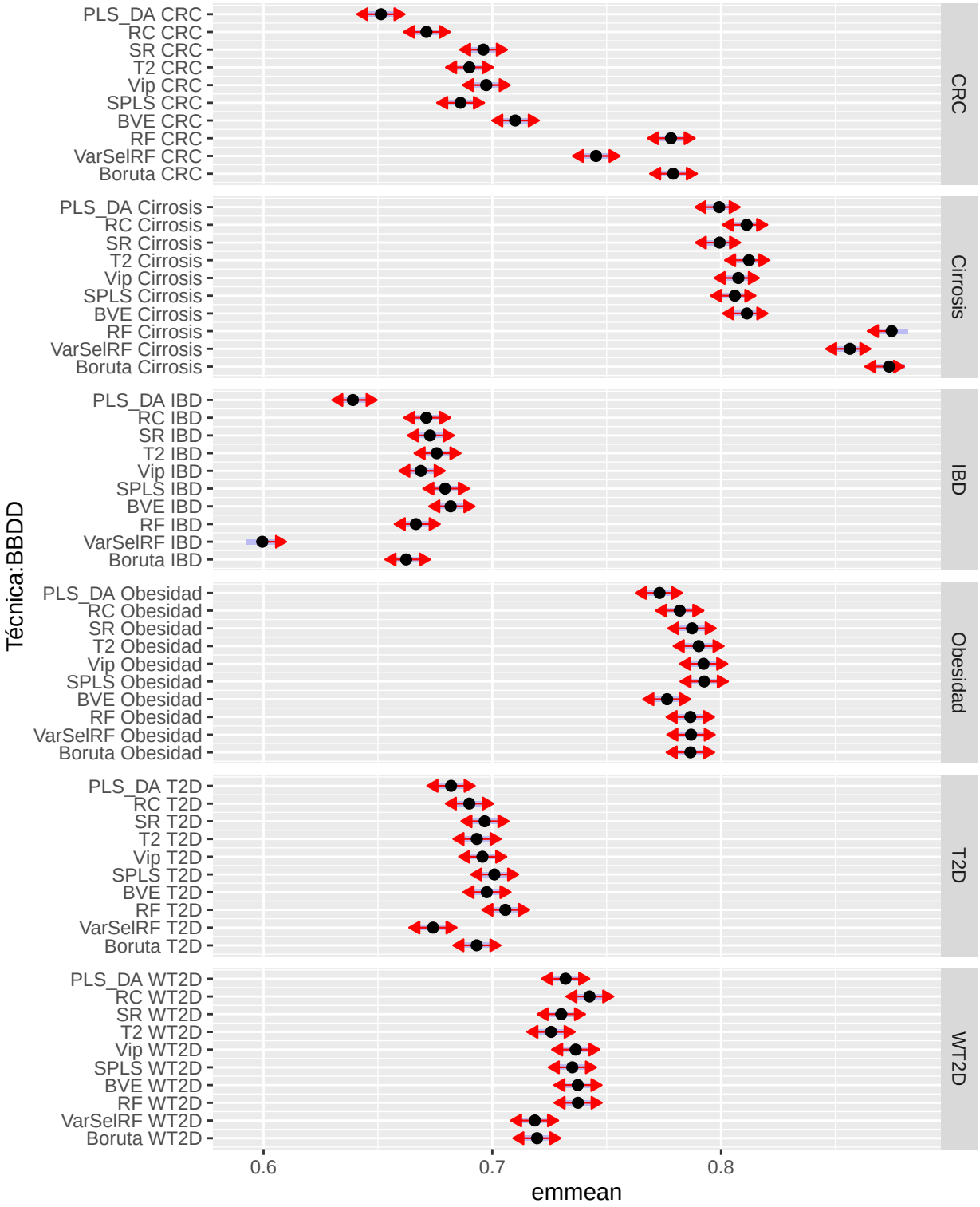

Figura A.1: Comparación MME del F1-score de la interacción técnica y BBDD

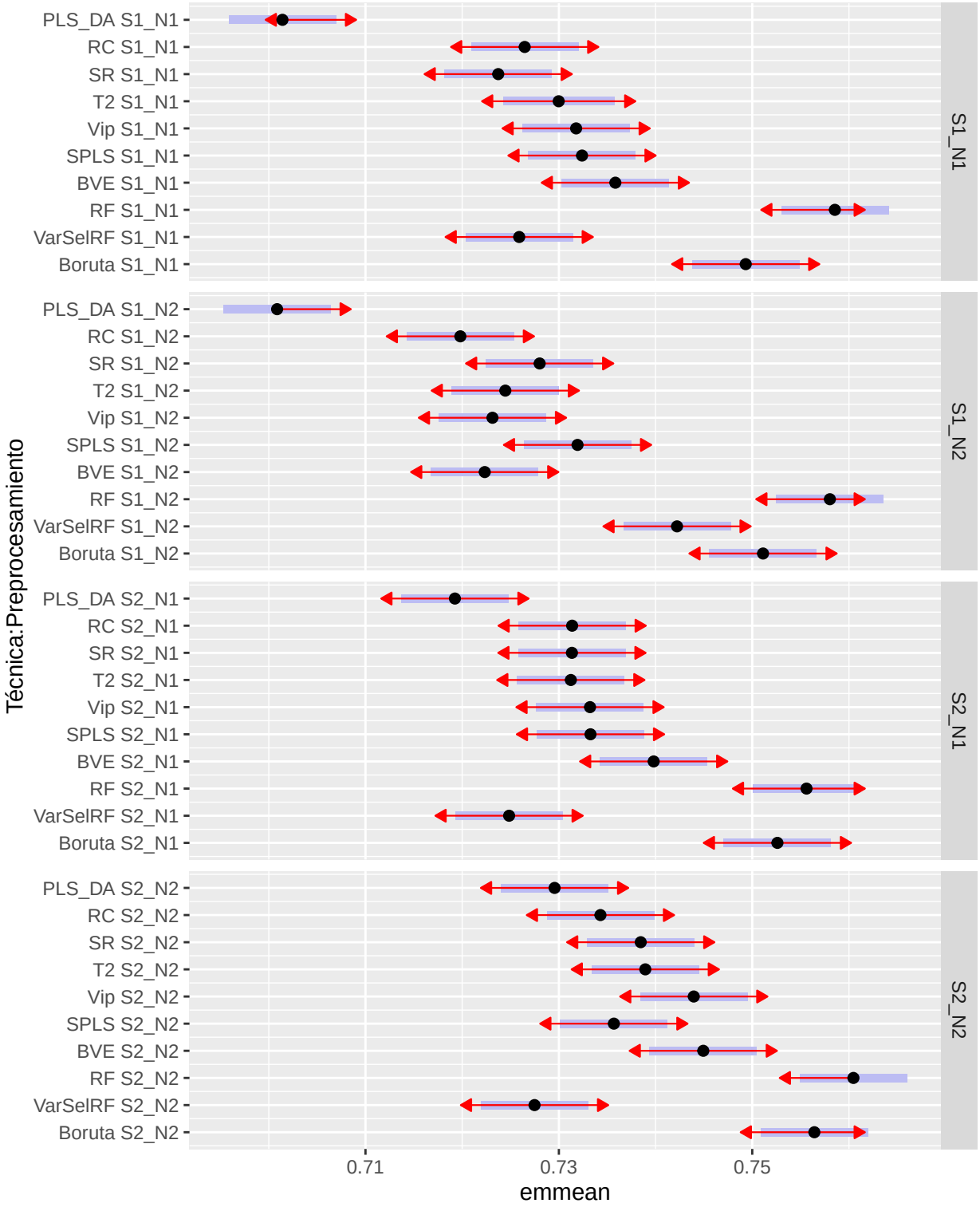

Figura A.2: Comparación MME del F1-score de la interacción técnica y el preprocesamiento de los datos

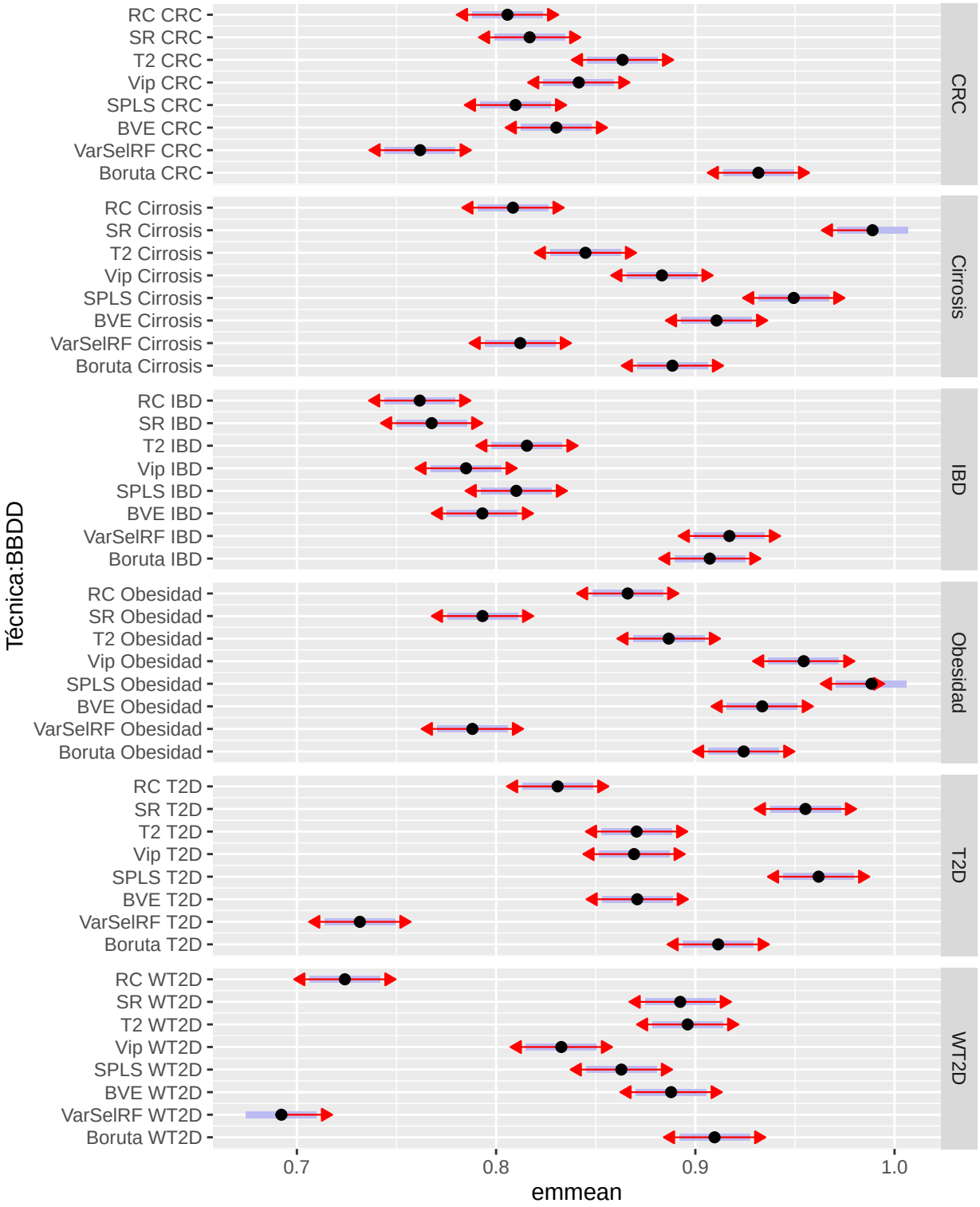

Figura A.3: Comparación MME de estabilidad de la interacción técnica y BBDD

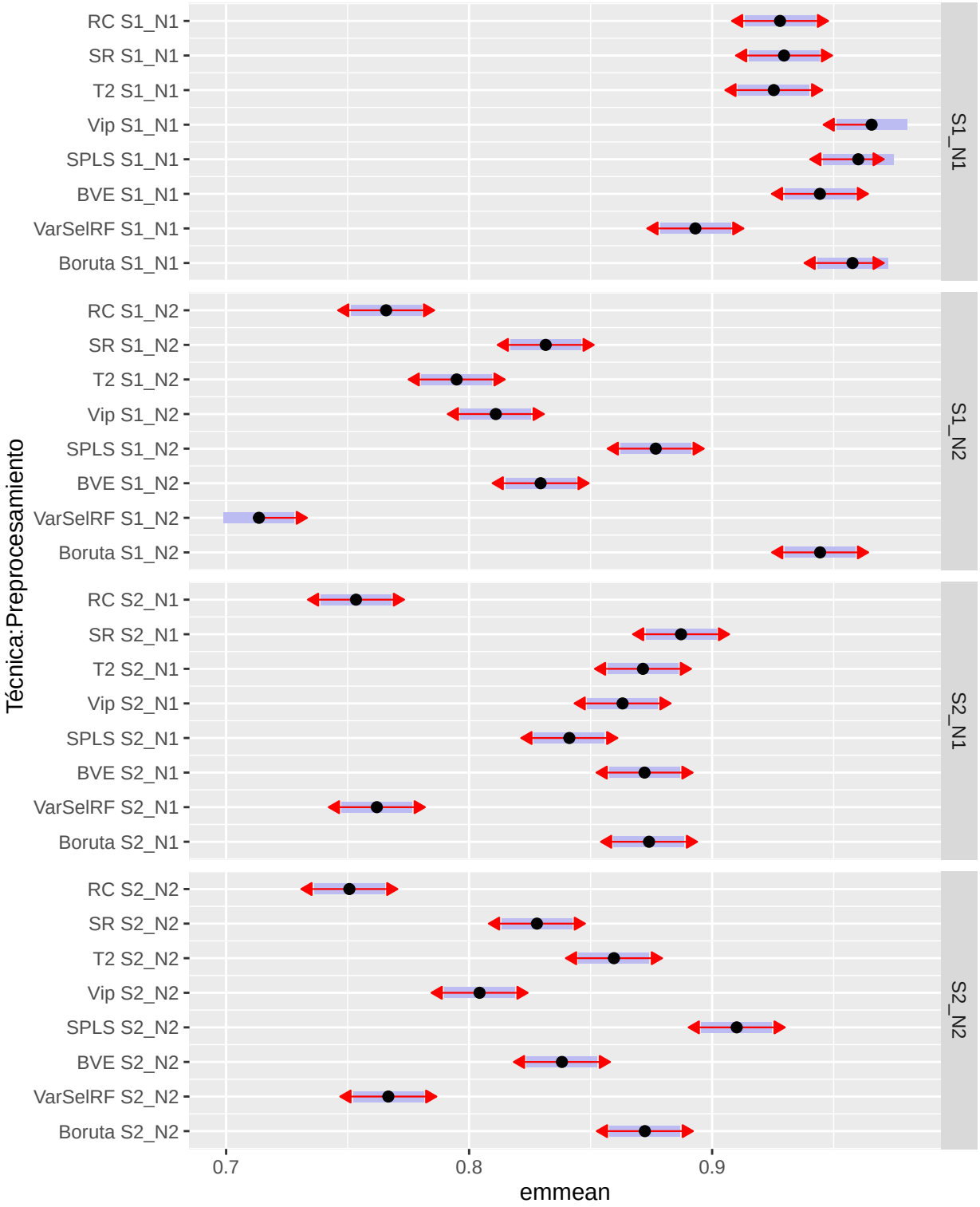

Figura A.4: Comparación MME de estabilidad de la interacción técnica y preprocesamiento de los datos

A.5. Número de Variables Seleccionadas

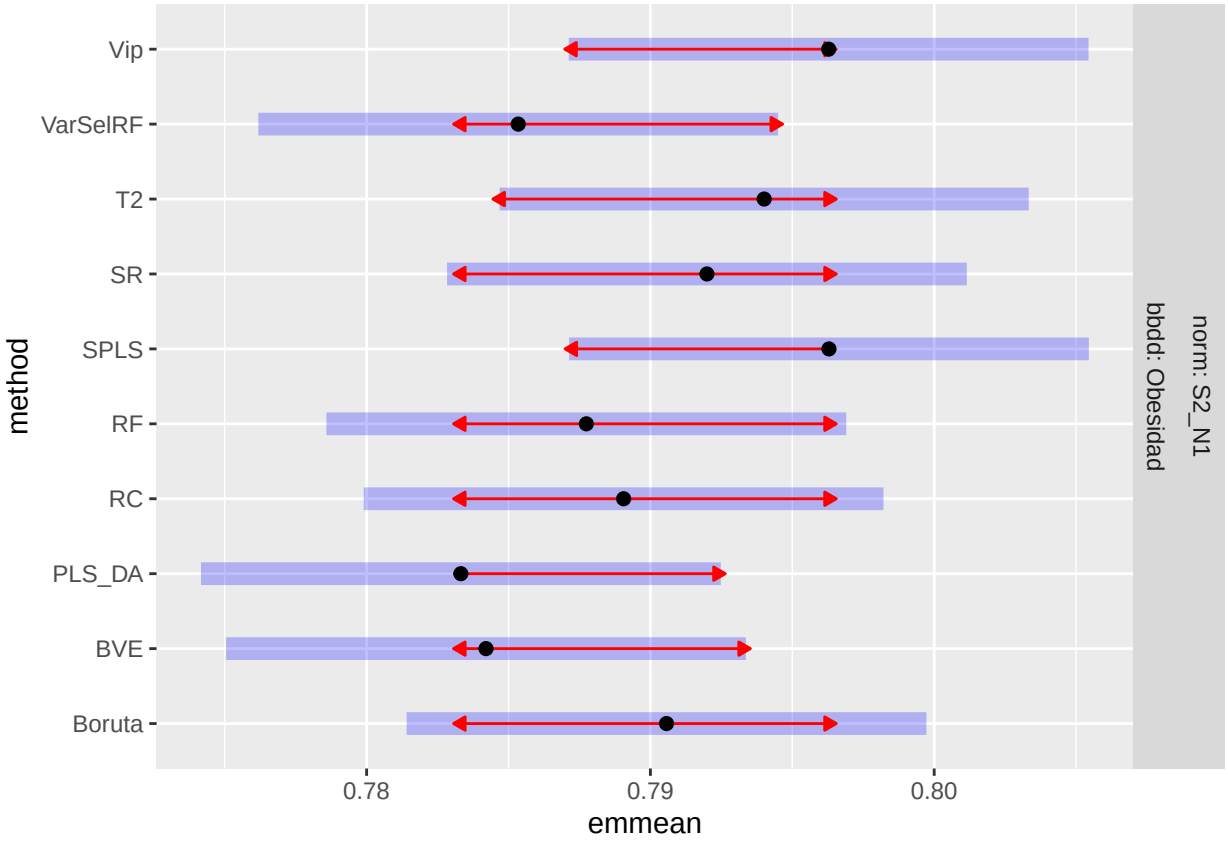

Figura A.5: Comparación MME de F1-Score de la técnicas para la BBDD Obesidad con el preprocesamiento de datos S2-N1

Tabla A.5: Nº Variables seleccionadas por técnica

Técnica	Mín.	Q25	Mediana	Q75	Máx.
Boruta	4	9	10	12	15
SPLS	2	3	4	5	9
Vip	2	3	4	4	5

A.6. Código

El código desarrollado se deja en un repositorio de Github:
`https://github.com/Samuelb1992/Codigo_TFM.git`

Bibliografía

[1] E. Dekaboruah, M. V. Suryavanshi, D. Chettri, y A. K. Verma, «Human Microbiome: An Academic Update on Human Body Site Specific Surveillance and Its Possible Role», *Arch Microbiol*, vol. 202, n.º 8, pp. 2147-2167, oct. 2020, doi: 10.1007/s00203-020-01931-x.

[2] R. E. Ley, P. J. Turnbaugh, S. Klein, y J. I. Gordon, «Human Gut Microbes Associated with Obesity», *Nature*, vol. 444, n.º 7122, 7122, pp. 1022-1023, dic. 2006, doi: 10.1038/4441022a.

[3] L. Wen *et al.*, «Innate Immunity and Intestinal Microbiota in the Development of Type 1 Diabetes», *Nature*, vol. 455, n.º 7216, 7216, pp. 1109-1113, oct. 2008, doi: 10.1038/nature07336.

[4] Q. Liu, Z. P. Duan, D. K. Ha, S. Bengmark, J. Kurtovic, y S. M. Riordan, «Synbiotic Modulation of Gut Flora: Effect on Minimal Hepatic Encephalopathy in Patients with Cirrhosis», *Hepatology*, vol. 39, n.º 5, pp. 1441-1449, 2004, doi: 10.1002/hep.20194.

[5] R. B. Sartor, «Microbial Influences in Inflammatory Bowel Diseases», *Gastroenterology*, vol. 134, n.º 2, pp. 577-594, feb. 2008, doi: 10.1053/j.gastro.2007.11.059.

[6] Z. Jie *et al.*, «The Gut Microbiome in Atherosclerotic Cardiovascular Disease», *Nat Commun*, vol. 8, n.º 1, 1, p. 845, oct. 2017, doi: 10.1038/s41467-017-00900-1.

[7] B. Wang, M. Yao, L. Lv, Z. Ling, y L. Li, «The Human Microbiota in Health and Disease», *Engineering*, vol. 3, n.º 1, pp. 71-82, feb. 2017, doi: 10.1016/J.ENG.2017.01.008.

[8] K. R. Amato, «An Introduction to Microbiome Analysis for Human Biology Applications», *American Journal of Human Biology*, vol. 29, n.º 1, p. e22931, 2017, doi: 10.1002/ajhb.22931.

[9] G. M. Weinstock, «Genomic Approaches to Studying the Human Microbiota», *Nature*, vol. 489, n.º 7415, 7415, pp. 250-256, sep. 2012, doi: 10.1038/nature11553.

[10] E. Pasolli, D. T. Truong, F. Malik, L. Waldron, y N. Segata, «Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights», *PLoS Comput Biol*, vol. 12, n.º 7, p. e1004977, jul. 2016, doi: 10.1371/journal.pcbi.1004977.

[11] M. D. Robinson y A. Oshlack, «A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data», *Genome Biol*, vol. 11, n.º 3, p. R25, 2010, doi: 10.1186/gb-2010-11-3-r25.

[12] S. Anders y W. Huber, «Differential Expression Analysis for Sequence Count Data», *Genome Biol*, vol. 11, n.º 10, p. R106, 2010, doi: 10.1186/gb-2010-11-10-r106.

[13] J. H. Bullard, E. Purdom, K. D. Hansen, y S. Dudoit, «Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments», *BMC Bioinformatics*, vol. 11, n.º 1, p. 94, feb. 2010, doi: 10.1186/1471-2105-11-94.

[14] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, y B. Wold, «Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq», *Nat Methods*, vol. 5, n.º 7, pp. 621-628, jul. 2008, doi: 10.1038/nmeth.1226.

[15] J. Aitchison, «The Statistical Analysis of Compositional Data», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 44, n.º 2, pp. 139-177, 1982, Disponible en: `https://www.jstor.org/stable/2345821`

[16] M. Badri, Z. D. Kurtz, C. L. Müller, y R. Bonneau, «Normalization Methods for Microbial Abundance Data Strongly Affect Correlation Estimates», 31 de agosto de 2018. `https://www.biorxiv.org/content/10.1101/406264v1` (accedido 15 de octubre de 2023).

[17] S. Wold, K. Esbensen, y P. Geladi, «Principal Component Analysis», *Chemometrics and Intelligent Laboratory Systems*, vol. 2, n.º 1, pp. 37-52, ago. 1987, doi: 10.1016/0169-7439(87)80084-9.

[18] M. J. Anderson y T. J. Willis, «Canonical Analysis of Principal Coordinates: A Useful Method of Constrained Ordination for Ecology», *Ecology*, vol. 84, n.º 2, pp. 511-525, 2003, doi: 10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2.

[19] D. Lee y H. S. Seung, «Algorithms for Non-negative Matrix Factorization», en *Advances in Neural Information Processing Systems*, 2000, vol. 13. Accedido: 15 de octubre de 2023. [En línea]. Disponible en: `https://proceedings.neurips.cc/paper_files/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html`

[20] L. Van der Maaten y G. Hinton, «Visualizing Data Using T-SNE.», *Journal of machine learning research*, vol. 9, n.º 11, 2008, Accedido: 19 de octubre de 2023. [En línea]. Disponible en: `https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl`

[21] J. I. E. Hoffman, «Logistic Regression», en *Biostatistics for Medical and Biomedical Practitioners*, 2015, pp. 601-611. doi: 10.1016/B978-0-12-802387-7.00033-0.

[22] P. A. Lachenbruch y M. Goldstein, «Discriminant Analysis», *Biometrics*, vol. 35, n.º 1, pp. 69-85, 1979, doi: 10.2307/2529937.

[23] G. Guo, H. Wang, D. Bell, Y. Bi, y K. Greer, «KNN Model-Based Approach in Classification», en *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, 2003, pp. 986-996. doi: 10.1007/978-3-540-39964-3_62.

[24] J. J. Werner *et al.*, «Impact of Training Sets on Classification of High-Throughput Bacterial 16s rRNA Gene Surveys», *ISME J*, vol. 6, n.º 1, 1, pp. 94-103, ene. 2012, doi: 10.1038/ismej.2011.82.

[25] C. Cortes y V. Vapnik, «Support-Vector Networks», *Mach Learn*, vol. 20, n.º 3, pp. 273-297, sep. 1995, doi: 10.1007/BF00994018.

[26] L. Breiman, «Random Forests», *Machine Learning*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.

[27] L. J. Marcos-Zambrano *et al.*, «Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment», *Front. Microbiol.*, vol. 12, p. 634511, feb. 2021, doi: 10.3389/fmicb.2021.634511.

[28] M. Barker y W. Rayens, «Partial Least Squares for Discrimination», *Journal of Chemometrics*, vol. 17, n.º 3, pp. 166-173, 2003, doi: 10.1002/cem.785.

[29] N. Romero y M. Camila, «Análisis de datos de microbioma para la predicción y caracterización de enfermedades», *Microbiome data analysis for disease prediction and characterization*, abr. 2023, Accedido: 9 de septiembre de 2023. [En línea]. Disponible en: `https://riunet.upv.es/handle/10251/192903`

[30] M. E. Ritchie *et al.*, «Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies», *Nucleic Acids Research*, vol. 43, n.º 7, p. e47, abr. 2015, doi: 10.1093/nar/gkv007.

[31] V. G. Tusher, R. Tibshirani, y G. Chu, «Significance Analysis of Microarrays Applied to the Ionizing Radiation Response», *Proceedings of the National Academy of Sciences*, vol. 98, n.º 9, pp. 5116-5121, abr. 2001, doi: 10.1073/pnas.091062498.

[32] M. B. Kursa y W. R. Rudnicki, «Feature Selection with the Boruta Package», *Journal of Statistical Software*, vol. 36, n.º 11, pp. 1-13, 2010, Disponible en: `https://doi.org/10.18637/jss.v036.i11`

[33] C. Ding y H. Peng, «MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA», 2005.

[34] S. Wold, M. Sjöström, y L. Eriksson, «PLS-regression: A Basic Tool of Chemometrics», *Chemometrics and Intelligent Laboratory Systems*, vol. 58, n.º 2, pp. 109-130, oct. 2001, doi: 10.1016/S0169-7439(01)00155-1.

[35] «A Wavelength Selection Method Based on Randomization Test for Near-Infrared Spectral Analysis | Elsevier Enhanced Reader». `https://reader.elsevier.com/reader/sd/pii/S0169743909000951?token=1B4C13BF95B9AFDFDD8D1D1FEF7D0F2ED0AB1120D59E9E030F4A1730BD4C3F3E149248937D916&originRegion=eu-west-1&originCreation=20221005181434` (accedido 5 de octubre de 2022).

[36] D. W. Aha y R. L. Bankert, «A Comparative Evaluation of Sequential Feature Selection Algorithms», en *Learning from Data: Artificial Intelligence and Statistics V*, D. Fisher y H.-J. Lenz, Eds. New York, NY: Springer, 1996, pp. 199-206. doi: 10.1007/978-1-4612-2404-4_19.

[37] L. Li, C. R. Weinberg, T. A. Darden, y L. G. Pedersen, «Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method», *Bioinformatics*, vol. 17, n.º 12, pp. 1131-1142, dic. 2001, doi: 10.1093/bioinformatics/17.12.1131.

[38] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, y I. Tsamardinos, «Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets», 10 de noviembre de 2016. `http://arxiv.org/abs/1611.03227` (accedido 14 de agosto de 2024).

[39] R. Tibshirani, «Regression Shrinkage and Selection Via the Lasso», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, n.º 1, pp. 267-288, ene. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[40] H. Zou y T. Hastie, «Regularization and Variable Selection Via the Elastic Net», *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, n.º 2, pp. 301-320, abr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[41] A. E. Hoerl y R. W. Kennard, «Ridge Regression: Applications to Nonorthogonal Problems».

[42] T. Mehmood, K. H. Liland, L. Snipen, y S. Sæbø, «A Review of Variable Selection Methods in Partial Least Squares Regression», *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62-69, ago. 2012, doi: 10.1016/j.chemolab.2012.07.010.

[43] T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K.-M. Myhr, y O. M. Kvalheim, «Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles», *Anal. Chem.*, vol. 81, n.º 7, pp. 2581-2590, abr. 2009, doi: 10.1021/ac802514y.

[44] O. M. Kvalheim, «Interpretation of Partial Least Squares Regression Models by Means of Target Projection and Selectivity Ratio Plots», *Journal of Chemometrics*, vol. 24, n.º 7–8, pp. 496-504, 2010, doi: 10.1002/cem.1289.

[45] Y. Saeys, I. Inza, y P. Larrañaga, «A Review of Feature Selection Techniques in Bioinformatics», *Bioinformatics*, vol. 23, n.º 19, pp. 2507-2517, oct. 2007, doi: 10.1093/bioinformatics/btm344.

[46] T. Mehmood, S. Sæbø, y K. H. Liland, «Comparison of Variable Selection Methods in Partial Least Squares Regression», *Journal of Chemometrics*, vol. 34, n.º 6, p. e3226, 2020, doi: 10.1002/cem.3226.

[47] T. Mehmood, «Hotelling T2 Based Variable Selection in Partial Least Squares Regression», *Chemometrics and Intelligent Laboratory Systems*, vol. 154, pp. 23-28, may 2016, doi: 10.1016/j.chemolab.2016.03.001.

[48] D. Chung y S. Keles, «Sparse Partial Least Squares Classification for High Dimensional Data», *Stat Appl Genet Mol Biol*, vol. 9, n.º 1, p. 17, mar. 2010, doi: 10.2202/1544-6115.1492.

[49] J. L. Speiser, M. E. Miller, J. Tooze, y E. Ip, «A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling», *Expert Systems with Applications*, vol. 134, pp. 93-101, nov. 2019, doi: 10.1016/j.eswa.2019.05.028.

[50] M. B. Kursa y W. R. Rudnicki, «Feature Selection with the **Boruta** Package», *J. Stat. Soft.*, vol. 36, n.º 11, 2010, doi: 10.18637/jss.v036.i11.

[51] R. D'ıaz-Uriarte y S. Alvarez de Andr'es, «Gene Selection and Classification of Microarray Data Using Random Forest», *BMC Bioinformatics*, vol. 7, 2006, doi: 10.1186/1471-2105-7-3.

[52] R. A. Irizarry y R. A. Irizarry, «Cross Validation», *Introduction to Data Science*, pp. 507-522, nov. 2019, doi: 10.1201/9780429341830-29.

[53] S. Bernard, L. Heutte, y S. Adam, «Influence of Hyperparameters on Random Forest Accuracy», en *Multiple Classifier Systems*, vol. 5519, J. A. Benediktsson, J. Kittler, y F. Roli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 171-180. doi: 10.1007/978-3-642-02326-2_18.

[54] P. Probst, M. N. Wright, y A. Boulestei, «Hyperparameters and Tuning Strategies for Random Forest», *WIREs Data Min & Knowl*, vol. 9, n.º 3, p. e1301, may 2019, doi: 10.1002/widm.1301.

[55] P. Probst y A.-L. Boulesteix, «To Tune or Not to Tune the Number of Trees in Random Forest?», 16 de mayo de 2017. `http://arxiv.org/abs/1705.05654` (accedido 23 de agosto de 2024).

[56] T. M. Oshiro, P. S. Perez, y J. A. Baranauskas, «How Many Trees in a Random Forest?»

[57] A. Kalousis, J. Prados, y M. Hilario, «Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces», *Knowl Inf Syst*, vol. 12, n.º 1, pp. 95-116, may 2007, doi: 10.1007/s10115-006-0040-8.

[58] A. Gałecki y T. Burzykowski, *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-3900-4.

[59] R. V. Lenth, *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. 2022. Disponible en: `https://CRAN.R-project.org/package=emmeans`

[60] M. Kuhn, *Caret: Classification and Regression Training*. 2022. Disponible en: `https://CRAN.R-project.org/package=caret`

[61] K. H. Liland, B.-H. Mevik, y R. Wehrens, *Pls: Partial Least Squares and Principal Component Regression*. 2022. Disponible en: `https://CRAN.R-project.org/package=pls`

[62] D. Chung, H. Chun, y S. Keles, *Spls: Sparse Partial Least Squares (SPLS) Regression and Classification*. 2019. Disponible en: `https://CRAN.R-project.org/package=spls`

[63] H. Bengtsson, *Parallelly: Enhancing the 'parallel' Package*. 2023. Disponible en: `https://CRAN.R-project.org/package=parallelly`

[64] R. Gaujoux, *doRNG: Generic Reproducible Parallel Backend for 'foreach' Loops*. 2020. Disponible en: `https://CRAN.R-project.org/package=doRNG`

[65] M. Corporation y S. Weston, *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. 2022. Disponible en: `https://CRAN.R-project.org/package=doParallel`

[66] A. Liaw y M. Wiener, «Classification and Regression by randomForest», *R News*, vol. 2, n.º 3, pp. 18-22, 2002, Disponible en: `https://CRAN.R-project.org/doc/Rnews/`

[67] R. Diaz-Uriarte, «GeneSrF and varSelRF: A Web-Based Tool and R Package for Gene Selection and Classification Using Random Forest», *BMC Bioinformatics*, vol. 8, 2007, doi: 10.1186/1471-2105-8-328.

[68] H. Wickham *et al.*, «Welcome to the tidyverse», *Journal of Open Source Software*, vol. 4, n.º 43, p. 1686, 2019, doi: 10.21105/joss.01686.

[69] D. Bates, M. Mächler, B. Bolker, y S. Walker, «Fitting Linear Mixed-Effects Models Using lme4», *Journal of Statistical Software*, vol. 67, n.º 1, pp. 1-48, 2015, doi: 10.18637/jss.v067.i01.

[70] M. Duan, Y. Wang, Q. Zhang, R. Zou, M. Guo, y H. Zheng, «Characteristics of Gut Microbiota in People with Obesity», *PLoS One*, vol. 16, n.º 8, p. e0255446, ago. 2021, doi: 10.1371/journal.pone.0255446.

[71] Z. XU, W. JIANG, W. HUANG, Y. LIN, F. K. L. CHAN, y S. C. NG, «Gut Microbiota in Patients with Obesity and Metabolic Disorders — a Systematic Review», *Genes & Nutrition*, vol. 17, n.º 1, p. 2, ene. 2022, doi: 10.1186/s12263-021-00703-6.

[72] Y. Lin *et al.*, «Combing Fecal Microbial Community Data to Identify Consistent Obesity-Specific Microbial Signatures and Shared Metabolic Pathways», *iScience*, vol. 26, n.º 4, p. 106476, abr. 2023, doi: 10.1016/j.isci.2023.106476.

[73] A. S. Meijnikman *et al.*, «Distinct Differences in Gut Microbial Composition and Functional Potential from Lean to Morbidly Obese Subjects», *Journal of Internal Medicine*, vol. 288, n.º 6, pp. 699-710, 2020, doi: 10.1111/joim.13137.

[74] L. Abenavoli *et al.*, «Gut Microbiota and Obesity: A Role for Probiotics», *Nutrients*, vol. 11, n.º 11, 11, p. 2690, nov. 2019, doi: 10.3390/nu11112690.

[75] C. Haro *et al.*, «The Gut Microbial Community in Metabolic Syndrome Patients Is Modified by Diet», *The Journal of Nutritional Biochemistry*, vol. 27, pp. 27-31, ene. 2016, doi: 10.1016/j.jnutbio.2015.08.011.

[76] B. A. Petriz *et al.*, «Exercise Induction of Gut Microbiota Modifications in Obese, Non-Obese and Hypertensive Rats», *BMC Genomics*, vol. 15, n.º 1, p. 511, jun. 2014, doi: 10.1186/1471-2164-15-511.

[77] H. Chen *et al.*, «Alternation of the Gut Microbiota in Metabolically Healthy Obesity: An Integrated Multiomics Analysis», *Front. Cell. Infect. Microbiol.*, vol. 12, nov. 2022, doi: 10.3389/fcimb.2022.1012028.

[78] L. Liu, «Investigation Of Optimal Culture Conditions And Development Of A Protective Delivery Method Of Eubacterium Siraeum As A Potential Probiotic», *Wayne State University Theses*, ene. 2017, Disponible en: `https://digitalcommons.wayne.edu/oa_theses/626`

[79] X. Hu *et al.*, «Integrative Metagenomic Analysis Reveals Distinct Gut Microbial Signatures Related to Obesity», *BMC Microbiology*, vol. 24, n.º 1, p. 119, abr. 2024, doi: 10.1186/s12866-024-03278-5.

[80] D. Zs'alig *et al.*, «A Review of the Relationship between Gut Microbiome and Obesity», *Applied Sciences*, vol. 13, n.º 1, 1, p. 610, ene. 2023, doi: 10.3390/app13010610.

[81] P. D. Cani, «Gut Microbiota and Obesity: Lessons from the Microbiome», *Briefings in Functional Genomics*, vol. 12, n.º 4, pp. 381-387, jul. 2013, doi: 10.1093/bfgp/elt014.

[82] A. L. Cunningham, J. W. Stephens, y D. A. Harris, «A Review on Gut Microbiota: A Central Factor in the Pathophysiology of Obesity», *Lipids in Health and Disease*, vol. 20, n.º 1, p. 65, jul. 2021, doi: 10.1186/s12944-021-01491-z.