UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Modelización multivariante de los distintos sustratos patológicos hepáticos con análisis multiparamétrico de resonancia magnética

Trabajo de Fin de Máster

Autor: Jordi Lluzar Martí

Directores: José Miguel Carot Sierra, David Martí Aguado

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

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

Universitat Politècnica de València

2020

Indice

1 Resumen

1.1 English

Chronic Liver Diseases (CLD) enlist one of the most worldwide prevalent diseases, affecting around 25% of the global population, with 2 billion dealing with obesity, 400 million with the burden of diabetes amongst others causing around 2 million deaths per year. These alarming numbers urge to improve medical services. Albeit biopsy remains as the *gold standard* in liver diagnosis, new techniques need to be considered and researched to reduce costs and improve patient satisfaction. Biopsy implies some negative consequences such as low reproducibility, great variability between medical experts, highly invasive, costly, and sampling issues. Magnetic Resonance Imaging is one of the candidates to overcome the biopsy as the *gold standard* due to its high spatial resolution, great contrast sensitivity, lack of ionizing radiation and compatibility with contrast agents. We conducted a research using multivariate and machine learning techniques of all the pathological substrates: fibrosis, inflammation, steatosis, and iron levels including variable selection and multiparametric analysis. Multivariate analysis was used to understand the response variable relationship and to detect some observations that needed to be double checked, as sample classification remains partly inconsistent. Additionally, it is essential to observe separation between disease stages. Results show clear correlation between magnetic resonance with continuous and categorical variables of histological biopsy samples; notwithstanding, there are few improvements that needs to be assessed to change the traditional *gold standard*.

Keywords: Multivariate Analysis, Classification, Regression, Machine learning, liver pathological substrates, magnetic resonance.

1.2 Español

Las Enfermedades Hepáticas Crónicas (EPC) encabezan en la lista de enfermedades más prevalentes en todo el mundo, afectando alrededor del 25% de la población mundial, con 2 mil millones de personas con obesidad, 400 millones con diabetes, entre otras patologías, causando alrededor de 2 millones de muertes por año. Estas cifras alarmantes instan a mejorar los servicios médicos. Aunque la biopsia sigue siendo el referente en el diagnóstico de hígado, se deben considerar e investigar nuevas técnicas para reducir los costes y mejorar la satisfacción del paciente. La biopsia implica algunas consecuencias negativas como baja reproducibilidad, alta variabilidad dependiente del submuestreo y el observador, ser altamente invasiva y cara. La Resonancia Magnética es una de las candidatas a superar la biopsia como método de referencia por su alta resolución, gran sensibilidad al contraste, ausencia de radiaciones ionizantes y compatibilidad con los agentes de contraste. Se ha realizado una investigación utilizando técnicas de análisis multivariante y *machine learning* de todos los sustratos patológicos: fibrosis, inflamación, esteatosis y niveles de hierro, incluida la selección de variables y el análisis multiparamétrico. Se utilizó el análisis multivariante para comprender la relación variables respuesta y detección de observaciones que debían comprobar su clasificación histológica, debido a la inconsistencia en la clasificación. Además, es fundamental observar la separación entre las etapas de la enfermedad. Los resultados muestran una clara correlación entre la resonancia magnética con variables continuas y categóricas de muestras de biopsias histológicas, no obstante, se debe seguir investigando para implementar mejoras en los métodos de análisis de imagen para obtener un mejor refinamiento, y así poder cambiar el método de referencia diagnóstico.

Palabras clave: Análisis multivariante, clasificación, regresión, aprendizaje automático, sustratos patológicos hepáticos, resonancia magnética.

1.3 Valencià

Les Malalties Hepàtiques Cròniques (MHC) encapçalen a la llista de enfermetats més prevalents a tot el món, afectant al voltant del 25% de la població mundial, amb 2 mil milions de persones amb obesitat, 400 milions amb diabetis, entre altres patologies, causant al voltant de 2 milions de morts a l'any. Aquestes xifres alarmants insten a millorar els serveis mèdics. Tot i que la biòpsia segueix sent el referent en el diagnòstic de fetge, s'han de considerar i investigar noves tècniques per reduir els costos i millorar la satisfacció al pacient. La biòpsia implica algunes conseqüències negatives

com baixa reproductibilitat, alta variabilitat depenent del submostreig i l'observador, ser altament invasiva i cara. La Ressonància Magnètica és una de les candidates a superar la biòpsia com a mètode de referència per la seva alta resolució, gran sensibilitat al contrast, absència de radiacions ionitzants i compatibilitat amb els agents de contrast. S'ha realitzat una investigació mitjançant tècniques d'anàlisi multivariante i *machine learning* de tots els substrats patològics: fibrosi, inflamació, esteatosi i nivells de ferro, inclosa la selecció de variables i l'anàlisi multiparamètric. Es va utilitzar l'anàlisi multivariant per comprendre la relació de variables resposta i detecció d'observacions que havien de comprovar la seva classificació histològica, a causa de la inconsistència en la classificació. A més, és fonamental observar la separació entre les etapes de la malaltia. Els resultats mostren una clara correlació entre la ressonància magnètica amb variables contínues i categòriques de mostres de biòpsies histològiques, però, s'ha de seguir investigant per implementar millores en els mètodes d'anàlisi d'imatge per obtenir un millor refinament, i així poder canviar el mètode de referència diagnòstic.

Paraules clau: Anàlisi multivariant, classificació, regressió, aprenentatge automàtic, substrats patològics hepàtics, ressonància magnètica.

2 Introducción

Las enfermedades hepáticas crónicas (EPC) constituyen un problema mundial con una demanda creciente de atención médica. Inicialmente puede originarse ciertas enfermedades difusas hepáticas (EDH) que se relacionan con un amplio espectro de etiologías como infecciones de virus de la hepatitis, un consumo de alcohol y otras sustancias nocivas, el hígado graso, enfermedades metabólicas, enfermedades autoinmunes, etc. (Kershenobich y Weissbrod, 2003). La actividad inflamatoria y la Fibrosis derivada son características histopatológicas comunes y relevantes en la EPC. Los pacientes con inflamación y Fibrosis hepática en estadios iniciales pueden ser asintomáticos, generando una urgencia en la rapidez y la fiabilidad de la detección de estas alteraciones tisulares para iniciar un tratamiento lo antes posible en caso de ser necesario (Fattovich et al., 1997).

Las EHC se han convertido en un problema de salud debido al aumento en la mortalidad (2 millones de muertes, que representan el 3,5% de las muertes totales a nivel global) y a la carga médica asistencial que representa la cirrosis (Asrani, Devarbhavi, Eaton, y Kamath, 2019). La gravedad de la lesión debe medirse con precisión para establecer el pronóstico de cada paciente, identificado los de mayor riesgo de complicaciones que requieren seguimiento y evaluar las posibilidades terapéuticas. (Friedman, 2003). Existe una relación lineal entre la Fibrosis y la mortalidad que indica la importancia de la diagnosis y la terapia precoz en pacientes con una hepatopatía crónica (Molloy et al., 2011).

Tradicionalmente, la biopsia hepática se ha considerado el *gold standard* para su diagnóstico y la clasificación de su nivel gravedad. La evaluación histológica establecida se basa en sistemas de puntuación semicuantitativos basados en criterios bien definidos, aunque interpretativos relacionados con la gravedad de la EPC subyacente. Esta evaluación categórica subjetiva se basa en cambios arquitectónicos y celulares, pero no en alteraciones patológicas continuas numéricas (Chevallier, Guerret, Chossegros, Gerard, y Grimaud, 1994). Por lo tanto, los informes de biopsia estándar son propensos a una gran variabilidad, lo que limita una evaluación crítica de las características histológicas en el entorno de la investigación y reduce la reproducibilidad de los resultados. Por otra parte, la biopsia hepática tiene otras limitaciones bien conocidas, ya que la técnica es un procedimiento costoso e invasivo, no aceptado bien por los pacientes y que no se puede usar libremente para el seguimiento. Además, es propenso a errores de submuestreo y sufre de variabilidad por el interobservador en la puntuación patológica (Ratziu et al., 2005). Más recientemente, se ha propuesto el uso del análisis de imagen digital asistido por computadora (DIA) de las secciones histológicas obtenidas en diferentes etiologías de EPC para una cuantificación más precisa de algunas características histológicas como la esteatosis y la cantidad de hierro hepático (Rifai et al., 2011).

En las últimas décadas, se han evaluado los métodos de imagen no invasivos en un intento de desarrollar biomarcadores de imagen como una práctica clínica alternativa a la biopsia hepática para la detección temprana, la evaluación precisa y el monitoreo factible de estos pacientes (Grigorescu, 2006). De entre los distintos métodos propuestos, la resonancia magnética (RM) ofrece una opción más fiable y completa para la evaluación del daño hepático. No obstante, existe la posibilidad de realizar un estudio conjunto.

En este trabajo se hará uso de biomarcadores no invasivos séricos (de análisis sanguíneo), elastográficos de transición (marcadores no invasivos) y los de imagen (resonancia magnética).

El objetivo de este trabajo es utilizar técnicas estadísticas multivariantes de clasificación y predicción para realizar estimaciones de los diferentes sustratos patológicos hepáticos (Fibrosis hepática, inflamación, esteatosis y hierro hepático) utilizando datos procedentes de imágenes de RM multiparamétricas (ADC, D, D *, f, DDC, α) y otras covariables médicas, revisando críticamente la relevancia clínica de las técnicas de imagen para estas evaluaciones.

Se propone comparar la eficacia en la clasificación de algunas técnicas multivariantes clásicas como la regresión logística y otros modelos de regresión con algunas técnicas de las calificadas como *Machine Learning* como SVM, *Random Forest* y otras. Se ensayarán distintos pretratamientos de los datos basados en técnicas de reducción de la dimensión. La evaluación de los métodos de clasificación y predicción se realizará con curvas ROC y PRC.

3 Objetivos

3.1 General

- Desarrollar modelos estadísticos multivariantes que permitan relacionar las variables de las imágenes de resonancia magnética con la clasificación anatomopatológica obtenida por biopsia de los distintos sustratos patológicos para mejorar la calidad de vida de los pacientes que sufran dichas enfermedades hepáticas, cambiando así el modelo diagnóstico de referencia.

3.2 Específicos

- Analizar la capacidad predictiva de los datos de análisis de imagen de resonancia magnética mediante técnicas estadísticas multivariantes y observar el comportamiento de las distintas variables y pacientes.
- Utilizar técnicas de *machine learning* para comparar los resultados de predicción.
- Relacionar, mediante las técnicas descritas anteriormente, las variables obtenidas por DIA con:
    - El grado de enfermedad de los distintos sustratos patológicos para el caso de las variables categóricas.
    - El nivel cuantitativo de los distintos sustratos patológicos para el caso de las variables continuas.
- Relacionar los niveles cuantitativos de los distintos sustratos patológicos con el grado categórico de cada uno para observar cómo se distribuyen los niveles cuantitativos según las distintas clases.
- Aplicar estrategias de selección de variables y optimización de hiperparámetros.
- Identificar, en caso de existir, biomarcadores que estén relacionados con los distintos sustratos patológicos.
- Evaluar los modelos mediante validación cruzada, curvas ROC, PRC y el método de Obuchowski.

4 Objeto de estudio

4.1 Descripción del estudio

Este TFM abarca la última de las etapas de un estudio que abarca el diseño, la planificación, la obtención de muestras (analíticas, de resonancia magnética y otras), el procesamiento de las muestras y los análisis estadísticos (Ilustración 1).

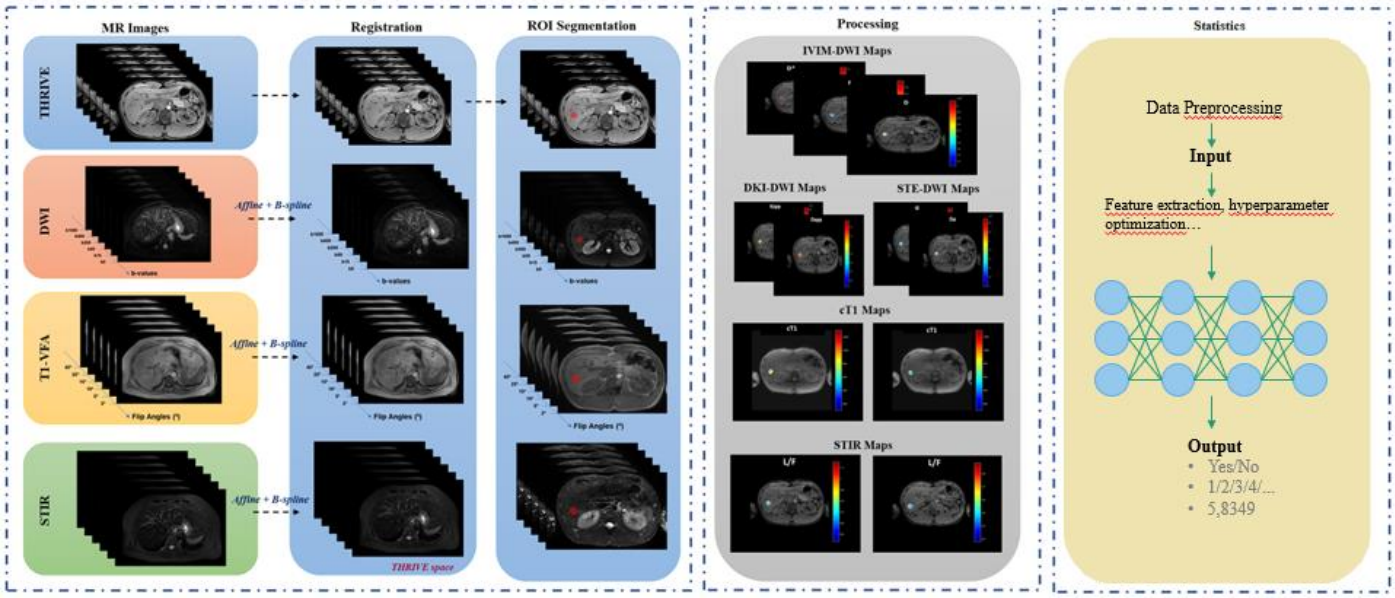


Ilustración 1 Diseño del estudio completo que comprende la parte de obtención de imágenes, su análisis a partir de las distintas fases y finalmente el estudio estadístico. Fuente: Elaboración propia.

El archivo de datos utilizado para el análisis estadístico forma parte de un estudio prospectivo multicéntrico de pacientes con hepatopatías crónicas difusas de los centros Hospital Clínico Universitario de Valencia (HCUV) y del Hospital Universitari i Politècnic La Fe (HUPLF). Pese a que a fecha de este TFM la base de datos se haya en estado incompleto, en total se incluirán 150 sujetos de estudio de manera prospectiva, todos con biopsia hepática independientemente de la hepatopatía crónica. Se tratará de ir incluyendo sujetos con el mayor espectro posible de la enfermedad (pacientes con estadios iniciales y con enfermedad avanzada). Los criterios generales de inclusión aceptan a mayores de 18 años con hepatopatía crónica que firmen el consentimiento informado y otorga una biopsia hepática con muestra histológica de calidad. Como criterios generales de exclusión se rechazan aquellos pacientes con contraindicaciones para la realización de la RM y/o presencia de enfermedad neoplásica como hepatocarcinoma o metástasis. El estudio ha sido aprobado por el Comité Ético de Investigación Biomédica de los hospitales HCUV y HUPLF, siguiendo todas las regulaciones del tratado de Helsinki.

En el diseño de estudio los pacientes se sometieron a una biopsia hepática como práctica clínica habitual (aguja de 16G con 20mm de muestra), tras firmar un consentimiento informado. Se insiste en el criterio de exclusión de eliminar del estudio a los pacientes que en la biopsia presentaron tejido tumoral o metástasis. Se realizaron otras técnicas analíticas (extracción y análisis de sangre, ecografías, etc.) para la recogida de variables clínicas y analíticas antes de la cuarta semana tras la biopsia. Finalmente se realizó el registro de las imágenes de resonancia magnética con las frecuencias MECSE para grasa y hierro, STIR-IVIM para la inflamación y T1W-mFA para Fibrosis, entre otras.

4.2 Obtención de Imágenes de Resonancia Magnética

En los siguientes apartados se explicará de manera simplificada el modo en el que se han obtenido las imágenes de resonancia magnética.

4.2.1 Protocolo de Resonancia magnética

Para la obtención de las imágenes[1] de resonancia magnética se expondrá a continuación, de manera breve, el protocolo de adquisición y las secuencias utilizadas para la obtención de los biomarcadores de imagen:

- IVIM-DWI[2]: Difusión con múltiples valores b (0,15,50,200,400,1000 s/mm^2).
- T1-mFA: Eco de gradiente con múltiples ángulos de inclinación (FA o *Flip Angles* en inglés) para calcular los tiempos de relajación T1 (FAs: 2,5,10,15,25, 45º).
- MECSE: Eco de gradiente multi eco con desplazamiento químico y múltiples tiempos de eco (TE) para cuantificar grasa y hierro (TEs: 0.9, 1.5, 2.2, 2.9, 3.5, 4.2, 4.9, 5.5, 6.2, 6.8, 7.5, 8.2 ms).
- STIR: *Short Time Inversion Recovery* para el edema (TI = 230 ms).
- THRIVE: 3D eco de gradiente de alta resolución potenciada en T1.

La principal ventaja de la obtención de imágenes por resonancia magnética (IRM) es que tienen características multidimensionales porqué existen diferentes procedimientos para generar imágenes (Drozdowicz, Bernasconi, Reyes, Saba, y Simón, 2005).

4.2.2 Preprocesado de Imagen digital

Tras la adquisición de las imágenes de RM en los pacientes, y previo al análisis de las diferentes imágenes RM, todas las secuencias fueron previamente preprocesadas para eliminar ruido y asegurar la coherencia entre los vóxeles adquiridos en las distintas series. Se emplearon filtros para ruido (non-local means) y de eliminación de inhomogeneidad de campo mediante filtros N4 (Kervrann, Boulanger, y Coupé, 2007).

Después de este proceso, se realizó una segmentación donde fueron seleccionadas tres regiones de interés (ROI) de 16mm de diámetro de las imágenes THRIVE, para cada paciente. Este proceso es especialmente útil para acortar el tiempo de procesado de las imágenes, de manera que se ignoran, durante el análisis, regiones que no son de interés, es decir, regiones periféricas. La segmentación es importante en aplicaciones como extracción de características anatómicas, medición volumétrica, compresión de datos, etc., ya que focaliza los métodos y el tiempo de computación la región de interés (Dolan y Jones, 2008).

[1] La calibración de la RM fue realizada a través de un fantoma de 15 viales cuyo contenido estaba realizado por diferentes concentraciones de grasa, hierro y colágeno; abarcando el rango de concentraciones que un ser humano puede contener en el hígado. Este procedimiento permite corregir posibles sesgos derivados de la adquisición o de la máquina, permitiendo una mayor reproducibilidad de los resultados.

[2] DWI significa *Diffusion-weighted imaging*.

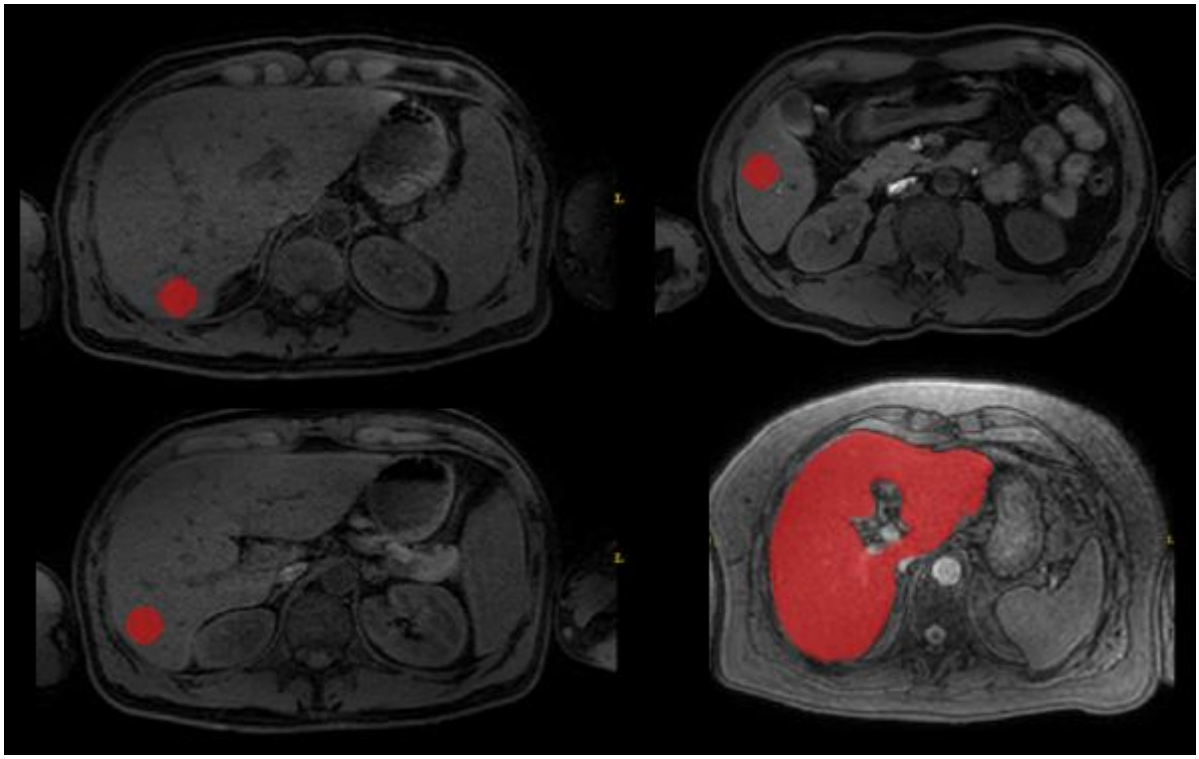


Ilustración 2 La segmentación ofrece la posibilidad de discriminar la región de interés del resto de los tejidos. En la imagen aparecen cuatro imágenes de un paciente con distintas regiones segmentadas posibles. Fuente: Archivo del Hospital Clínico Universitario de Valencia.

Después de ello, las imágenes IVIM-DWI, T1-mFA, MECSE y STIR fueron registradas inicialmente entre la propia secuencia (registro intra-secuencia) y posteriormente conjuntamente sobre las imágenes THRIVE (registro inter-secuencial) para garantizar así la coherencia espacial dentro de un espacio de referencia común (THRIVE). Los métodos de registro empleados fueron métodos rígidos y no rígidos (Dawant, 2002).

El registro es una técnica empleada en el procesado digital de imágenes que permite mediante transformadas matemáticas situar imágenes de diferentes adquisiciones en un mismo espacio, de manera que al trabajar con distintos pacientes o diferentes imágenes de un paciente estamos trabajando con la misma ROI, en la que los vóxeles de estudio son equivalentes. Con estas técnicas se puede corregir posibles movimientos del paciente que se hayan producido durante su adquisición, incluyendo la respiración constante, facilitando así la segmentación de las regiones de interés (Audette, Ferrie, y Peters, 2000). En la Ilustración 2 se pueden apreciar las distintas posibilidades de realizar un registro parcial o completo, según la región de interés.

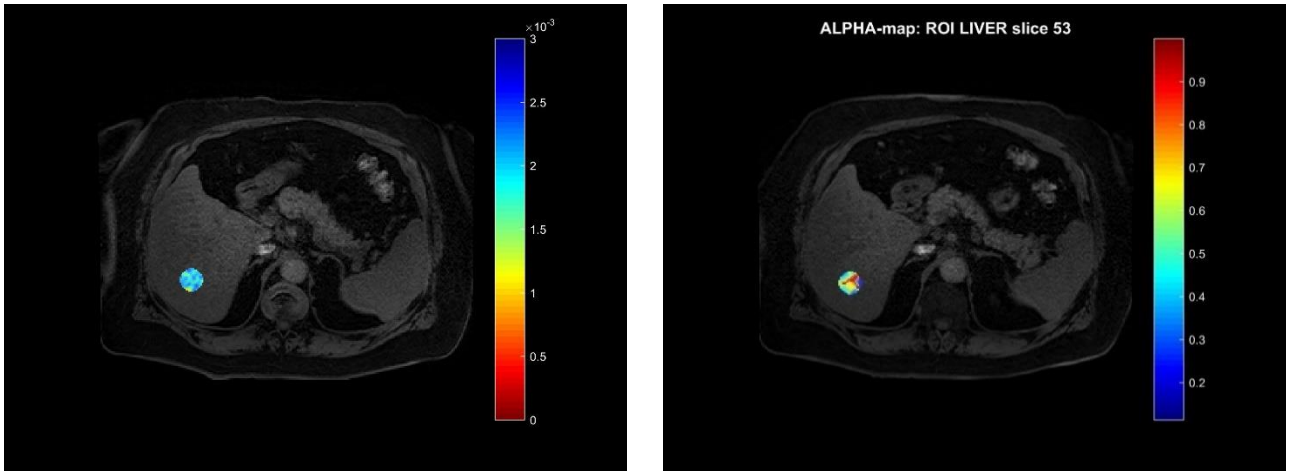


Ilustración 3 Imágenes obtenidas en la fase de segmentación facilitadas por el HUPLF. A la izquierda la secuencia ADC_2b y a la derecha ALPHA_STR. En ellas puede apreciarse el mapeo de cada variable en zona NASH, donde es esperable una mayor concentración de grasa y hierro. Fuente: Archivo del Hospital Clínico Universitario de Valencia.

En la fase de análisis se utilizan diferentes modelos matemáticos aplicados a las diferentes intensidades de las imágenes o a la evolución de esta a lo largo de un parámetro (tiempo de eco, diferentes valores b, adquisición con diferentes valores de FA, etc.) lo que permiten extraer diferentes características de la imagen relacionadas con diferentes procesos fisiológicos relacionado con las patologías bajo estudio tales como la esteatosis, la acumulación de hierro, la Fibrosis y la inflamación (Hatta et al., 2010). En la Ilustración 3 se aprecia una intensidad en los vóxeles distinta para las variables ADC con valor 2b y ALPHA.

Así para las imágenes T1-mFA se calculó el tiempo de relajación T1 (en milisegundos) aplicando un ajuste de curvas, vóxel a vóxel, a las señales formadas con los valores de intensidad respecto a los diferentes ángulos de inclinación.

Se realizó un proceso similar aplicado a las imágenes IVIM-DWI y MECSE. En estos casos el ajuste se realizó sobre las señales formadas con los valores de intensidad de las diferentes adquisiciones respecto a los diferentes valores b y a tiempo de eco, respectivamente. En estos casos el objetivo era sacar información respecto a la difusividad del tejido (IVIM-DWI) y valores de grasa y de hierro (MECSE) de las regiones seleccionadas.

Finalmente, para las adquisiciones de STIR, se midió la relación de la intensidad de la señal de hígado con respecto a la grasa (Donato, França, Candelária, y Caseiro-Alves, 2017).

4.3 Anatomía Patológica

Para la obtención cuantitativa y clasificación de los niveles de los distintos sustratos patológicos se ha realizado el siguiente procedimiento:

1. En la fase de biopsia se ha utilizado una aguja semiautomática de 16G de dos pasos con 20mm de muestra y 11 espacios porta. La biopsia se ha realizado en el lóbulo hepático derecho (segmento V-VI), se ha fijado en formol e incrustado en parafina.
2. Las tinciones histológicas para la detección de los distintos sustratos patológicos han sido:
    - Rojo sirio para Fibrosis.
    - IHQ adipofilina para grasa.
    - Tinción Perls para el hierro.
    - IHQ CD45 para la inflamación.
3. Análisis semicuantitativo: *NAS-CRN score system*. El Comité de Patologías de la *NASH Clinical Research Network* ha diseñado y validado un sistema de puntuación histológico que abarca un espectro completo de lesiones de la enfermedad del hígado graso no alcohólico (NAFLD en inglés). En la Ilustración 4 mostrada a

continuación se exponen los criterios para asignar una clasificación a las distintas muestras para los distintos sustratos patológicos (Kleiner et al., 2005).

Item	Definition	Score/Code
Steatosis		
Grade	Low- to medium-power evaluation of parenchymal involvement by steatosis	
	<5%	0
	5%-33%	1
	>33%-66%	2
	>66%	3
Location	Predominant distribution pattern	
	Zone 3	0
	Zone 1	1
	Azonal	2
	Panacinar	3
Microvesicular steatosis*	Contiguous patches	
	Not present	0
	Present	1
Fibrosis		
Stage		
	None	0
	Perisinusoidal or periportal	1
	Mild, zone 3, perisinusoidal	1A
	Moderate, zone 3, perisinusoidal	1B
	Portal/periportal	1C
	Perisinusoidal and portal/periportal	2
	Bridging fibrosis	3
	Cirrhosis	4
Inflammation		
Lobular inflammation	Overall assessment of all inflammatory foci	
	No foci	0
	<2 foci per 200× field	1
	2-4 foci per 200× field	2
	>4 foci per 200× field	3
Microgranulomas	Small aggregates of macrophages	
	Absent	0
	Present	1
Large lipogranulomas	Usually in portal areas or adjacent to central veins	
	Absent	0
	Present	1
Portal inflammation	Assessed from low magnification	
	None to minimal	0
	Greater than minimal	1
Liver cell injury		
Ballooning*		
	None	0
	Few balloon cells	1
	Many cells/prominent ballooning	2
Acidophil bodies		
	None to rare†	0
	Many	1
Pigmented macrophages		
	None to rare†	0
	Many	1
Megamitochondria*		
	None to rare†	0
	Many	1
Other findings		
Mallory's hyaline	Visible on routine stains	
	None to rare†	0
	Many	1
Glycogenated nuclei	Contiguous patches	
	None to rare†	0
	Many	1

Ilustración 4 Sistema de puntuación NAS-CRN para clasificar los distintos sustratos hepatopatológicos a partir de la histología de las muestras. Fuente: Kleiner et al., (2005).

4. Análisis cuantitativo: La cuantificación de los distintos sustratos patológicos se ha realizado a partir de un avanzado escáner (*iScan HT*). Después se han procesado las imágenes, incluyendo un proceso de normalización para poder obtener la distribución de los porcentajes de los distintos sustratos patológicos en las distintas muestras. Dichas imágenes se han analizado mediante MATLAB para obtener el porcentaje de los valores de la CPA (relacionado con Fibrosis), de CD45 (relacionado con inflamación), de FPA (relacionada con la grasa) y de hierro (Martí Aguado et al., 2020).

5 Materiales y métodos

5.1 Descripción de la base de datos

La base de datos disponible está constituida por $179\ filas\ x\ 155\ columnas$ con multitud de datos faltantes. Con el propósito de asegurar la confidencialidad de los pacientes se ha procedido a retirar el número de la seguridad social asociado a cada sujeto de manera previa a los análisis estadísticos. Además, para facilitar la trazabilidad de cada sujeto por el interés médico también se ha incluido el Hospital en el que se han realizado las pruebas.

A los pacientes se les han asignado un identificador llamado "NT" que consiste en una secuencia numérica del 1 al 179. Los NT se utilizarán de referencia para conservar la trazabilidad de cada paciente durante los análisis estadísticos.

El motivo de que existan más de 150 filas, que es la cifra que se pretende alcanzar en el estudio, es debido a que hay pacientes que han decidido salir del mismo o que inicialmente se había comprometido a asistir y después no se han presentado. Existe una columna llamada "Causa de no realización" donde encontramos diversas causas como la pérdida del estudio por falta del bloque histológico, artefactos que aparecen en las imágenes de resonancia, gente que ya no quiere seguir y contraindicaciones entre otras. Las causas más comunes de la salida de un estudio son, no obstante, la inasistencia por la crisis sanitaria del Covid-19, obesidad y claustrofobia. Otra columna llamada "Comentarios AP" detalla algunas incidencias en relación con la Anatomía patológica para casos particulares y en algunos casos se decide eliminar caso por el estado de las muestras.

El resto de las variables que podemos encontrar son las que el equipo médico ha considerado que podía tratarse de potenciales biomarcadores relacionados con las distintas hepatopatías. En ellas podemos encontrar información relativa al historial médico, análisis sérico o de sangre, análisis elastográfico de transición (biomarcadores mediante tecnología no invasiva), los resultados analíticos de IRM y la clasificación realizada por biopsia hepática mediante ARFI. En ellas podemos encontrar:

5.1.1 Variables Explicativas o matriz de Datos X

- **Historial médico:** Edad (numérica), Sexo (Hombre/Mujer), Índice de Masa Corporal o IMC (numérica), Síndrome metabólico (Sí/No), Hipertensión Arterial o HTA(Si/No), Diabetes Mellitus o DM (Sí/No), Dislipemia o DL (Sí/No), Hipotiroidismo (Sí/No), Fumador (Sí/No).
- **Análisis sérico:** Plaquetas (numérica), Hormona Estimulante de la Tiroides o TSH (numérica), Hormona T4 o Tiroxina o T4L (numérica), Glucosa (numérica), Leucocitos (numérica), Neutrófilos (numérica), Linfocitos (numérica), Monocitos (numérica), Eosinófilos (numérica), Ratio neutrófilos/linfocitos, Hematíes, RDW ancho distribución eritrocitaria (numérica), Ratio RDW/plaquetas (numérica), Ferritina (numérica), Aspartato aminotransferasa o GOT (numérica), Transaminasa glutámico pirúvica o GPT (numérica), Ratio GOT/GPT (numérica), Gamma-glutamil transferasa o GGT (numérica), Ratio GGT/plaquetas (numérica), Triglicéridos o TG (numérica), Colesterol (numérica), LDL (numérica), HDL (numérica), Ratio TG/glucosa, Ratio TG/HDL, Ratio LDL/HDL, Etiología (CBP/HT/DILI/HT/OH/NASH/VHC), Framingham Score I (numérica), Framingham Score II (numérica), Riesgo de enfermedad cardiovascular aterosclerótica o ASCVD (numérica),
- **Marcadores no invasivos de Fibrosis mediante variables analíticas:** APRI (numérica), FIB-4 (numérica), Índice de BARD (0/1/2/3/4), *Non-Alcohol Fatty Liver Disease Fibrosis Score* o NAFLD FS (numérica), Índice BAAT (0/1/2/3/4), Albumina-bilirubina o ALBI (numérica), Índice Forns (numérica).

- **Marcadores no invasivos de Esteatosis mediante variables analíticas:** *Hepatic Steatosis Index* o HSI (numérica).
- **Análisis elastográfico de transición (no invasivo)[3]:** TE Kpa [4](numérica), TE CAP [5](numérica).
- **Análisis digital de imagen de resonancia magnética:** Las variables de resonancia magnética vienen con la media, mediana, desviación estándar, percentil 25 y percentil 75 obtenidos obtenidas tras la medición de intensidad de los vóxeles. A continuación, para simplificar se expondrán los grupos de variables para las distintas secuencias.
    - **Secuencia IVIM:** Coeficiente de Difusión (D), Coeficiente de pseudo-perfusión (D^*), fracción vascular (*f*), Coeficiente de Difusión Aparente (ADC) con valores 2b, ADC con valores 6b.
    - **Secuencia KURT:** Difusión aparente (Dapp)[6], Kurtosis (Kurt).
    - **Secuencia Stretch:** Coeficiente de difusión (Dalpha)[7], Índice de heterogeneidad del intravoxel (Alpha).
    - **Mapas:** Tiempo de relajación corregido al hierro (cT1), Tiempo de relajación (T1).
    - **Secuencia MECSE:** *Fat Fraction* (FF), *Water Fraction* (WF), tiempo de relajatividad transversal al agua. (R2W), y tiempo de relajatividad transversal en grasa (R2F).
    - **Secuencia STIR:** Ratio entre la intensidad del vóxel del tejido hepático y la grasa periférica (RATIO).

5.1.2 Variables Respuesta o matriz de Datos Y

A continuación, se detallan las variables respuesta utilizadas en el TFM de los distintos sustratos patológicos con los niveles correspondientes que comprenden, así como de las variables categóricas y sus *endpoints*.

- **Variables de anatomía patológica:**
    - **Semicuantitativas por sustratos patológicos:**
        - **Fibrosis:** Score (0-4), Score dicotomizada (0-2 vs 3-4), ISHAK (0-6), ISHAK dicotomizada (0-4 vs 5-6).
        - **Inflamación:** Lobular (0-3), Lobular dicotomizada (0-1 vs 2-3), Portal (0-1), Batts Ludwig (0-4), Batts Ludwig dicotomizada (0-2 vs 3-4), Inflamación avanzada (0-1)[8].
        - **Grasa:** Esteatosis (0-3), Esteatosis dicotomizada (0-1 vs 2-3).
        - **Hierro:** Scheuer (0-4), Scheuer dicotomizada (0 vs 1-4), Deugniers (0-32).
    - **Cuantitativas:**
        - *Collagen Proporcional Area* (CPA).
        - *Cluster of Differentiation 45* (CD45).
        - *Fat Proporcional Area* (FPA).
        - Porcentaje de Hierro (Hierro).

[3] En el artículo de A.Cequera & Méndez se detalla una descripción elaborada de la obtención y el significado de dichas variables en las que no se entrará en detalle en este trabajo para no incluir información irrelevante en la materia de Análisis de Datos.
[4] Marcador no invasivo de la Fibrosis.
[5] Marcador no invasivo de la esteatosis.
[6] La difusión aparente (Dapp) se interpreta igual que la difusión (D), la diferencia es la aplicación de distintos modelos a las imágenes DWI (IVIM y Kurtosis, respectivamente).
[7] En la misma línea de lo comentado en el punto 2, el coeficiente de difusión Dalpha es el resultado tras la aplicación del modelo Stretch a las imágenes DWI.
[8] El criterio de la inflamación avanzada contempla asignar como grado 1 si en cualesquiera de los otros tipos de inflamación dicotomizadas se ha asignado un grado 1.

5.1.2.1 Explicación médica acerca de los endpoints

El criterio seguido para establecer los puntos de corte que en la dicotomización de las variables categóricas es meramente clínico. Por normal general, los pacientes dicotomizados como grado 1 presentan un nivel avanzado de la enfermedad y los síntomas son graves, en el grado 0 presentan cierto grado de enfermedad.

En el caso de la Fibrosis, la literatura relata un incremento de la mortalidad a partir del grado 2 en comparación con el grupo control, pero de una manera más atenuada en comparación con clases superiores. Los análisis de supervivencia predicen las clases 3 y 4 mucho mejor el aumento de la futura mortalidad que las clases inferiores, sobre todo al ajustar covariatas, incluyendo la edad (Hagström, et al., 2017). Para la esteatosis, se dicotomiza con grados 0, 1 *versus* 2 y 3 en base al riesgo de mortalidad creciente entre las distintas clases (Nasr, Fredrikson, Ekstedt, y Kechagias, 2020).

Para el hierro con la variable Scheuer un 0 vs 1-4 significa que el grado 0 no presenta hierro en el tejido hepático, y con 1-4 agrupamos los distintos niveles de hierro existentes. En la literatura se puede observar que los niveles de hierro se asocian con el resto de las enfermedades avanzadas (Fibrosis score 3-4, inflamación Lobular 2-3, Portal 1, y esteatosis 2-3), debido en parte a que el estrés oxidativo provoca cambios celulares que inducen enfermedad (Nelson et al., 2011). No obstante, cuanto mayor es el grado de Fibrosis (clase 4) menor es la relación con el resto de las enfermedades debido a que la acumulación de colágeno hepático impide la acumulación del resto de componentes asociados a las diversas hepatopatologías (Martí Aguado et al., 2020).

5.1.3 Procedimiento general en la modelización de los sustratos patológicos

Para el presente trabajo se ha seguido un enfoque de trabajo dedicado a optimizar las predicciones de cada hepatopatía estudiada. Para ello se han utilizado modelos tanto paramétricos como no paramétricos para no caer en suposiciones sobre la distribución de las variables. Lo que se expone a continuación se aplica tanto a variables continuas como categóricas, para la metodología seguida.

Tal y como se expresa en la Ilustración 5, tras el análisis exploratorio y el pretratamiento de datos se han utilizado distintos modelos estadísticos con la finalidad de estudiar la relación entre las variables explicativas y las variables respuesta desde distintos puntos de vista estadísticos. El enfoque utilizado ha sido el de tratar de eliminar el máximo ruido a través de la selección de variables que no estén estrechamente relacionadas con la variable respuesta. En algunos casos, estos métodos de selección de variables están asociados a características específicas de cada modelo (como la media de reducción en Gini para *Random Forest*). En otros casos se ha empleado métodos más generales basados en importancia o en la selección por otros modelos (como utilizar SVM tras la selección de variables por un algoritmo de eliminación recursiva de variables basado en *Random Forest*). En cualquier caso, siempre se ha utilizado como método de referencia el modelo estadístico con todas las variables. El próximo paso es la selección de parámetros o de hiperparámetros, necesario para poder maximizar la eficacia estadística en cada modelo, según sea paramétrico o no paramétrico, respectivamente.

Los modelos estadísticos, los métodos de selección de variables, y los métodos de optimización de hiperparámetros utilizados se expondrán más adelante en el apartado de Estado del Arte.

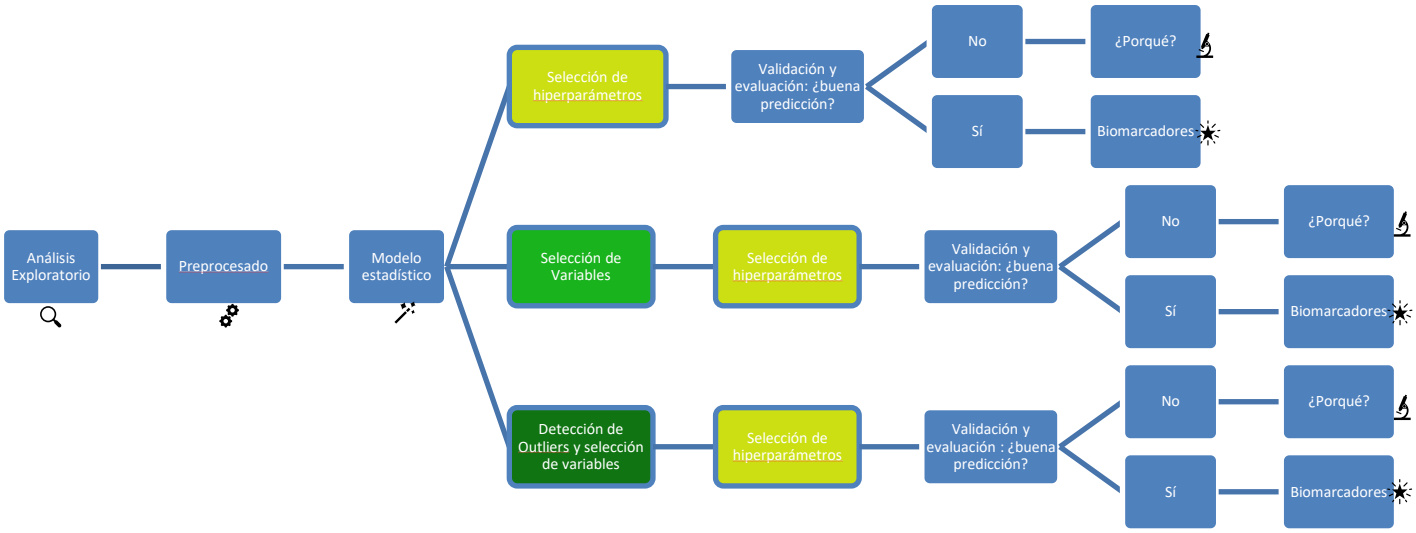


Ilustración 5 Metodología general aplicada del TFM en la sección de relacionar las variables explicativas X con las variables respuesta Y. De manera general se han realizado pruebas con todas las variables y con la selección de estas, y el caso de los métodos basados en estructuras latentes, a excepción de LS-SVM, se ha realizado un análisis de residuos para detectar las observaciones anómalas. Fuente: Elaboración propia.

5.1.4 Comparación entre variables categóricas y continuas

La relación entre los tipos de variables respuesta (categóricas y continuas) es especialmente útil para observar la distribución de los valores continuos en las diversas categorías de los sustratos patológicos. El objeto de estudio modelizar los distintos sustratos patológicos, y tras la biopsia realizada, se han clasificado las distintas muestras de anatomía patológica, y a su vez, se han cuantificado por los niveles de dicha enfermedad, previamente teñidas, tal y como se ha detallado anteriormente. El método escogido para evaluar las diferencias entre las distintas clases ha sido el de Kruskal-Wallis (KW). Como existen diversas variables multiclase, se ha querido ir más allá comentando las diferencias entre pares mediante la prueba de Wilcoxon.

5.2 Software utilizado

El archivo de datos se obtuvo en formato ".xlsx" de **Microsoft Excel** (Microsoft Corporation, 2019). Se ha utilizado Excel para la visualización de la base de datos y se ha manipulado para incluir nuevas variables dicotómicas a partir de las variables preexistentes.

Para el análisis estadístico se ha recurrido al software libre **R** (R Core Team, 2020) a través del programa **RStudio** (RStudio Team, 2020). Los paquetes o librerías utilizados para la realización de este trabajo son los siguientes:

- **readxl** (Wickham y Bryan, 2019): Permite cargar los archivos Excel para su posterior utilización.
- **dplyr** (Wickham, François, Henry, y Müller, 2020): Ofrece una extensa flexibilidad en la manipulación de datos a partir de una gramática particular, facilitando el trabajo y ahorrando líneas de código.
- **car** (Fox y Weisberg, 2019): Aporta diversas herramientas relacionadas con la regresión. En este trabajo se ha utilizado para realizar la prueba de homocedasticidad (test de Levene).

- **e1071** (Meyer, Dimitriadou, Hornik, Weingessel, y Leisch, 2019): Esta librería es muy polivalente para el uso de diversos modelos y pruebas estadísticas. Se ha empleado para el uso de SVM y optimizar diversos parámetros mediante *Grid search*.
- **randomForest** (Liaw y Wiener, 2002): Permite utilizar *Random Forest* y aplicar los cálculos y visualización de importancias.
- **VSURF** (Genuer, Poggi, y Tuleau-Malot, 2019): Utiliza una selección de variables usando *Random Forest*s.
- **mixOmics** (Rohart F. , Gautier, Singh, y Cao, 2017): Se compone de una gran cantidad de herramientas multivariantes que implementan diversas estrategias. Se ha utilizado para llevar a cabo un sPLS-DA y un sPLS.
- **mdatools** (Kucheryavskiy, 2020): Esta librería también se compone de múltiples herramientas multivariantes. Se ha usado como complemento de la anterior para llevar a cabo la detección de observaciones anómalas o extremas, como paso previo a la explotación del modelo.
- **caret** (Kuhn, 2020): El paquete caret implementa muchos otros modelos del ecosistema R para adaptarse a un abanico realmente elevado de posibilidades. Su uso se ha limitado a la selección recursiva de variables basado en *Random Forest*.
- **glmnetcr** (Archer y Williams, 2012): Permite el uso de la Regresión Logística Ordinal con penalización Lasso para variables categóricas multiclase.
- **glmnet** (Friedman, Hastie, y Tibshirani, 2010): Es un paquete equivalente al anterior con múltiples posibilidades de selección (penalización ridge, lasso y elasticnet) para el caso de variables categóricas binarias o dicotómicas.
- **rminer** (Cortez, 2020): Este paquete se ha utilizado por la implementación de la selección de variables mediante un algoritmo heurístico basado en SVM.
- **DMwR** (Torgo, 2010): Permite el uso de *Tree Regression* y *Tree Classification*.
- **kernlab** (Karatzoglou, Smola, Hornik, y Zieleis, 2004): Este paquete implementa múltiples métodos estadístico. Se ha utilizado por su implementación de *Least Squares Support Vector Machine*.
- **ROCPR** (Grau, Grosse, y Keilwagen, 2015): En la evaluación de los modelos estadísticos compuesto por variables categóricas binarias se han utilizado las curvas ROC y PR, implementadas en esta librería. También aporta el cálculo de las áreas bajo la curva ROC y PR.
- **nonbinROC** (Nguyen, 2012): Esta librería ha sido útil para el cálculo de la probabilidad en la precisión del método de Obuchowski para variables no binarias o multiclase. La librería permite realizar el cálculo para variables categóricas multiclase ordinales y nominales, así como para continuas.
- **cutpointr** (Thiele y Hirschfeld, 2020): El uso de esta librería ha permitido calcular el estadístico J conocido como índice de Youden, además de la especificidad, sensibilidad y precisión óptimos.

5.3 Pretratamiento de Datos

5.3.1 Autoescalado

Centrar y escalar las variables es un proceso muy habitual en técnicas estadísticas multivariantes. En este TFM se ha decidido aplicar un centrado y escalado para todos los modelos estadísticos, es decir, haciendo que nuestros datos se distribuyan con media cero $\mu = 0$ y varianza unitaria $\sigma = 1$.

Mediante el centrado se consigue que el centro de coordenadas de las variables originales se sitúe en el centro de gravedad de la nube de puntos generada por los datos. Es decir, cambiamos la escala de la variable, aunque la distancia entre las puntuaciones de la variable permanece inalterable. Para centrar se resta una constante a todas las observaciones de cada variable, aunque lo más habitual es la media de cada variable. El objetivo del centrado es eliminar los efectos de escala que una variable pueda tener sobre las demás.

$$x_{ij}{}^{centrada} = x_{ij} - \bar{x}_j \quad \text{Ec.1}$$

$$\bar{x}_j = \frac{\sum_{i=1}^{I} x_{ij}}{I} \qquad \text{Ec.2}$$

La siguiente transformación aplicada es la estandarización de las variables. Esto se realiza dividiendo a las puntuaciones de las distintas variables por la desviación típica de cada variable. Si lo aplicamos a las variables ya centradas, quedarían centradas y reducidas o estandarizadas.

$$x_{ij}{}^{cent.+estand.} = \frac{x_{ij}{}^{centrado}}{\sigma_j} \qquad \text{Ec.3}$$

De esta manera los datos ahora son independientes a la escala utilizada y las variables tienen la misma media y dispersión, además de que los coeficientes de correlación permanecen inalterados a las variables sin tratar (Berg, Hoefsloot, Westerhuis, Smilde, y Werf, 2006). Para los análisis de datos es más fácil trabajar de esta manera puesto que se pueden comparar más fácilmente las variaciones.

5.3.2 Imputación de Datos Faltantes

Existen numerosos datos faltantes en la base de datos debido a que se haya en estado incompleto. La inexistencia de datos puede dificultar el análisis puesto que se trabaja con un tamaño de muestra menor al esperado y las pruebas de contraste de hipótesis tienen una potencia menor. Puede producirse además una disminución de la representatividad de la muestra (Medina y Galván, 2007).

La imputación consiste en estimar los valores ausentes en base a los valores existentes de las otras variables o casos de la muestra. Se puede aplicar a un conjunto de variables o algunas especialmente seleccionadas.

En este TFM no se van a utilizar métodos de imputación debido a que podríamos estar introduciendo un sesgo en las observaciones al no conocer la distribución real y algunos de los aspectos determinantes en la transformación de imágenes en datos. En el futuro se espera poder corregir este problema ampliando la base de datos hasta el fin del estudio.

5.3.3 Balanceo de clases

Pese a que se trata de incluir en el estudio a sujetos con un amplio espectro de la enfermedad, la posterior clasificación puede acabar resultando muy distinta y aparecen pocos sujetos con la enfermedad en estado avanzado, quizá se deba a que buscan un tratamiento médico cuando aparecen los síntomas de la enfermedad. En cualquier caso, en el archivo de datos existen diversas hepatopatías en las cuales la proporción de las clases de enfermedad no es homogénea.

Cuando utilizamos modelos estadísticos con datos muy desbalanceados normalmente las predicciones son peores debido a que el modelo se entrena con muchas clases de uno o varios tipos y poco de otros tipos. Como ejemplo simplificado, si tenemos las predicciones basadas en una clasificación binaria muy desbalanceada obtienen los mejores resultados si apuntamos hacia la clase con más casos en la predicción.

Uno de los algoritmos más comunes es el SMOTE (*Synthetic Minority Over-Sampling Technique*) que crea nuevas observaciones de manera no paramétrica a partir de los *k* vecinos más cercanos. El resultado final es un set de datos balanceado con la misma proporción (o una proporción similar, según la configuración del algoritmo) en el que se han añadido nuevas observaciones sintéticas (Douzas, 2018).

Balancear los datos para este TFM puede resultar interesante, puesto que en general, todas las variables explicativas categóricas están desbalanceadas. No obstante, se ha decidido prescindir de dicha estrategia debido a que existen

algunas variables categóricas con una única observación de una clase determinada. Así pues, como se tienen muy pocas observaciones de algunas clases, no conocemos cómo se distribuye la enfermedad en estado avanzado.

Finalmente, para no realizar ningún tipo de suposiciones, se procedió a realizar pruebas de varios modelos con clases balanceadas y no se produjo una mejora basada en el porcentaje de aciertos del clasificador.

5.4 Modelos estadísticos utilizados

A continuación, se explica el fundamento teórico de los modelos utilizados para reflejar la decisión de actuación sobre nuestros datos a partir de un procedimiento enfocado en obtener la mejor predicción y extraer la información explicativa de dichos modelos.

La mayoría de las técnicas se pueden utilizar para regresión y clasificación. Las técnicas de regresión utilizan variables respuesta continuas y la predicción también es un valor de tipo continuo. Las técnicas de clasificación utilizan variables cualitativas o discretas y las predicciones son una clase o una categoría. Todas las técnicas expuestas a continuación se pueden utilizar para regresión y clasificación excepto sPLS-DA y la Regresión logística ordinal (para variables originales y latentes) que son exclusivas de clasificación. De la misma manera, sPLS es una técnica exclusiva para regresión.

5.4.1 Modelos basados en estructuras latentes

5.4.1.1 Sparse Partial Least Squares Regression

El objetivo de PLS es hallar la máxima variación en los espacios de datos definidos como variables explicativas **X** y la variable respuesta **Y**, que a diferencia del PCA tan solo maximiza la varianza explicada del espacio de las **X**. En PLS (y PLS-DA), la máxima variación entre los dos espacios de datos se obtiene maximizando la covarianza entre **X** e **Y**. Dicha covarianza es equivalente al coeficiente de correlación cuando estandarizamos previamente los datos, por ello es importante normalizar los datos, además de centrarlos (Ec. 4). Con ello obtenemos las componentes o variables latentes cuyas direcciones vienen determinadas por el espacio con máxima covarianza (Geladi y Kowalski, 1986).

$$Cov(X,Y) = r(X,Y) \cdot s_X \cdot s_Y \qquad \text{Ec.4}$$

Siendo *r(X, Y)* el coeficiente de correlación entre X e Y, y *s* la desviación típica de cada elemento X e Y.

Tras la obtención de la primera dirección principal en las **X**, la segunda es obligatoriamente ortogonal a la primera, y así lo será sucesivamente hasta llegar al mínimo (*n*-1, *p)*, siendo *n* el número de observaciones y *p* el número de variables. No obstante, en las **Y** el principio de ortogonalidad no se debe cumplir necesariamente (Dunn, 2020).

Las coordenadas definidas por el hiperplano de cada par de componentes principales se obtienen a través de los *scores*, y se nombrarán *scores* **t** para el espacio de las **X** y **u** para los *scores* del espacio de las **Y**. La relación lineal entre las coordenadas de las proyecciones de las **X** e **Y** se puede observar enfrentado gráficamente los *scores* $\mathbf{t_1}$ y $\mathbf{u_1}$, a través de su estructura interna, y observaremos que esta relación lineal irá perdiéndose de manera general en cada par de *scores*.

Los *loadings* **p** se corresponden a las direcciones en el espacio de las **X**. No obstante, en el modelo PLS se utilizan los pesos **w** asociados a la correlación entre las variables del espacio de las **X** y de las **Y**, con la particularidad de que en la primera dimensión los pesos $\mathbf{w_{1}}=\mathbf{w^*_1}$. A partir de la segunda dimensión, el vector $\mathbf{w_{a=2}}$ se calcula a partir de la matriz tras la deflación en **X**, por lo que la interpretación de estos *loadings* nos ofrece la relación entre los *scores* $\mathbf{t_2}$ y la matriz

$\mathbf{X_2}$ deflacionada. Esta deflación se realiza para eliminar variabilidad que ya ha sido explicado en $\mathbf{X_a}$ e $\mathbf{Y_a}$. De esta manera, para interpretar correctamente con la matriz de datos original utilizaríamos $\mathbf{w_a^*=w(p^Tw)^{-1}}$.

En el modelo sPLS se ha introducido la suposición de *sparsity* o escasez, que supone que solo un pequeño número de variables explicativas están relacionadas con la variable respuesta. De esta manera y para restar complejidad al modelo se aplica una penalización ℓ_1 en el vector de *loadings* asociado a la matriz **X**, reduciendo la contribución de estas variables irrelevantes en la formación del modelo (Chung, Chun, & Keles, 2012). La aproximación se basa en la Descomposición en Valores Singulares (SVD en inglés) del producto cruzado $M_h = X_h^T Y_h$. Los vectores singulares de la SVD, y para un número h de deflaciones o dimensiones elegidas en el modelo PLS, serían u_h y v_h. Se aplica la penalización ℓ_1 a los *loadings* en u_h y v_h. La optimización del problema de PLS se resuelve con la minimización de la regla de Frobenius entre el producto matricial y los vectores singulares (*loadings)*, tal y como vemos en la siguiente expresión:

$$\min_{u_h, v_h} \|M_h - u_h v_h'\|^2 F + P_{\lambda 1}(u_h) + P_{\lambda 2}(v_h) \qquad \text{Ec.5}$$

Donde $P_{\lambda 1}(u_h) = signo(u_h)(|(u_h)| - \lambda 1)_+$, y $P_{\lambda 2}(u_h) = signo(v_h)(|(v_h)| - \lambda 2)_+$ se aplican a cada componente escogida en el modelo en los vectores u_h y v_h y son los umbrales elegidos en las funciones que aproximan la función de penalización Lasso, aplicadas simultáneamente en ambos vectores de *loadings*. La Ecuación 5 se resuelve iterativamente y las matrices X_h y Y_h se deflacionan en cada iteración h (Lê Cao, Boitard, y Besse, 2011).

5.4.1.2 Sparse Partial Least Squares Discriminant Analysis

El análisis discriminante por mínimos cuadrados parciales (PLS-DA) es una particularización de la regresión por mínimos cuadrados parciales (PLS) con variables respuesta categóricas (Lee, Liong, y Jemain, 2018). Tanto es así, que se desarrollan tantos modelos PLS como clases, asignando cada vez una de estas clases de manera que posteriormente se puede determinar la pertenencia de una observación dependiendo de en qué modelo se ajuste mejor. Este método presenta muchas ventajas respecto al análisis discriminante clásico puesto que es eficaz frente a datos faltantes y multicolinealidad entre las variables explicativas. La estructura del modelo y sus elementos se representan en la siguiente figura:

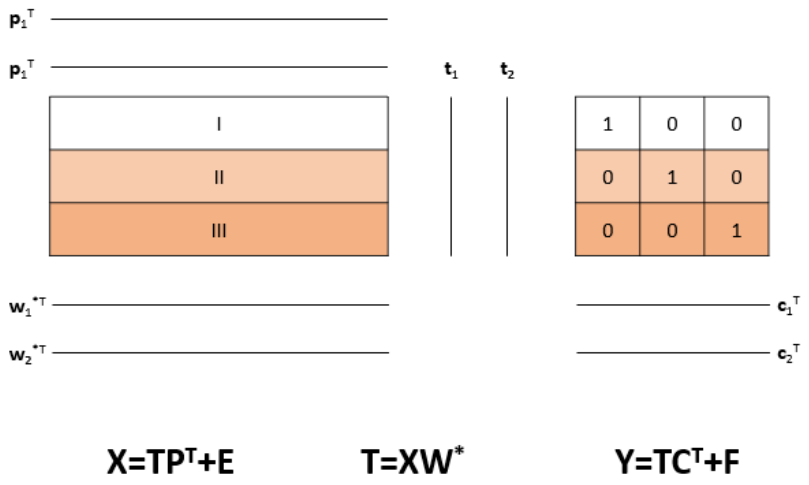


Ilustración 6 Figura del modelo de PLS-DA. Fuente: elaboración propia.

La predicción de un modelo PLS-DA es un valor que oscila entre cero y uno para cada clase, siendo cero la no pertenencia a la clase y uno la pertenencia a dicha clase. En la práctica se utiliza un umbral para la decisión, que normalmente es de 0,5. La probabilidad de cada predicción se obtiene aplicando la distribución normal a los valores predichos de **Y** a partir de PLS y después calculando la probabilidad de un valor dado de **Y** (Worley, Halouska, y Powers, 2013).

En el modelo sPLS-DA también se aplica el concepto de sparsity, explicado en el apartado anterior (Rohart, Gautier, Singh, y Lê Cao, 2017).

5.4.1.3 Least Squares Support Vector Machine

Least Squares Support Vector Machine o LS-SVM (Vieira, Neto, & Rodrigues, 2016) es un método basado en Support Vector Machine (SVM). El método usa un conjunto de ecuaciones lineales durante el proceso de entrenamiento del modelo mediante las variables explicativas, por lo que, al generalizar el modelo, se reduce el sobreajuste y la complejidad computacional. Para ello, solo se requiere la capacidad de poder calcular la inversa de una matriz. Por otro lado, el resultado se aleja de la simplicidad de poder reducir el número de variables, ya que todas las observaciones se utilizan como vectores soporte. El modelo es eficaz frente a un razonable número de datos faltantes.

Los modelos de LS-SVM tienen a calcular la función de pérdida con el criterio de los mínimos cuadrados, de manera que la norma del error cuadrático representa la función de pérdida. Desde un punto de vista de la optimización, se podría representar la función objetivo tal como:

$$min\ {}^{J}_{\boldsymbol{w},b,\xi}(\boldsymbol{w},\xi) = \frac{1}{2}\boldsymbol{w^T w} + \gamma\ \frac{1}{2}\sum_{i=1}^{n}\xi_i^2 \qquad \text{Ec.6}$$

con las restricciones de igualdad de:

$$y_i = w^T\emptyset(\boldsymbol{w_i}) + b + \xi_i^2, i = 1,\ldots n. \quad \text{Ec.7}$$

Y el modelo de LS-SVM se expresaría de la siguiente manera:

$$f(x) = \boldsymbol{w^T}\emptyset(x) + b \quad \text{Ec.8}$$

donde J es la función objetivo, $1/2\ w^T w$ representa la función de medida de planitud o llanura, ξ_i^2 es el error variable, γ es el factor de compensación entre la planitud del modelo y el error de entrenamiento (ecuación 6), y el mapeo no lineal $\emptyset$ representa la entrada de datos en un espacio con una dimensión elevada que se debe resolver por regresión lineal, b es el sesgo y **w** es el vector de *loadings* asociados a la relación entre el espacio de las variables explicativas y la variable respuesta (ecuación 7).

Las ecuaciones anteriores se pueden unificar de manera que:

$$L(\boldsymbol{w}, b, \xi, \alpha) = \frac{1}{2}\boldsymbol{w^T w} + \gamma\, \frac{1}{2}\sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i(\xi_i^2 - y_i + \boldsymbol{w^T}\emptyset(x_i) + b) \qquad \text{Ec.9}$$

Donde $\sum_{i=1}^{n} \alpha_i$ se refiere a los multiplicadores de Lagrange. Para resolver el problema, la función de Lagrange se optimiza diferenciando con respecto a **w**, b, α_i y ξ_i. Los resultados se pueden formular en un sistema lineal de ecuaciones para representar el problema de clasificación. El paso resultante sería ajustar la función Kernel, tal que:

$$K(x_i, x_j) = \emptyset(x_i)^T \cdot \emptyset(x_j) \qquad \text{Ec.10}$$

De manera que el modelo LS-SVM que se utiliza para la estimación de la función es:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_j, x_i) + b \qquad \text{Ec.11}$$

Existen diversos tipos de kernels: radial, sigmoidal, polinomial, etc. El más utilizado en la literatura es el radial (RBF). De esta manera, se requiere la elección libre de dos parámetros: el parámetro de coste particular a cada tipo de kernel tras su elección y el valor de regularización positiva γ. Por ejemplo, en la expresión del kernel RBF se debe elegir el valor de sigma.

$$K(x_j, x_i) = exp\{^{-\|x_j - x_i\|^2}/_{2\sigma^2}\} \qquad \text{Ec.12}$$

Así pues, valores muy pequeños de γ[9] o valores muy elevados de σ van a resultar en un mal ajuste del modelo (*underfitting* en inglés). De manera opuesta, se podría inducir un sobreajuste en el modelo (*overfitting* en ingles). El ajuste del Kernel se ha realizado a partir de los procesos gaussianos ya que en la literatura aparecen como métodos eficaces frente a otro tipo de técnicas como algoritmos genéticos (Afshin, Sadeghian, y Raahemifar, 2007).

5.4.2 Modelos de aprendizaje supervisado

5.4.2.1 Ordinal Logistic Regression with Lasso Penalization

En este trabajo se han utilizado variables categóricas de más de dos clases (multiclase) y tras su dicotomización se han convertido en binarias. Para ello se emplean los modelos de regresión logística binomial (Hosmer y Lemesbow, 1980) y regresión logística ordinal (Bender y Grouven, 1997). La regresión logística ordinal es una particularización de la regresión logística puesto que es un modelo logístico acumulativo o de *odds* proporcionales.

[9] Para no confundir la formulación con el modelo de SVM, en LS-SVM γ equivale el parámetro de regularización, llamado C en SVM. El parámetro del Kernel radial σ de LS-SVM equivale a γ en SVM.

En la regresión logística se pueden utilizar varias funciones para calcular la probabilidad de que ocurra o no un evento suceda entre 0 y 1. La más empleada es la *logit*, puesto que suaviza el efecto de los regresores sobre la respuesta. Intuitivamente se puede afirmar que la probabilidad de que suceda un evento no es lineal con las variables de estudio.

El modelo *logit* se formularía de la siguiente manera:

$$g(p_j) = logit(p_j) = ln\frac{p}{1-p} = \overrightarrow{x_j}\,\overrightarrow{\beta} \quad ; \quad \frac{p}{1-p} = e^{\overrightarrow{x_j}\,\overrightarrow{\beta}} \qquad \text{Ec.13}$$

Como podemos ver, el modelo se puede transformar en un modelo lineal sobre la variable *logit*. Dicha variable representa en escala logarítmica la diferencia de probabilidades de pertenecer a cada categoría. El cociente $p/1-p$ se le conoce como *odd*. Para la interpretación de parámetros se recurre al cociente entre *odds*, llamado odd-ratio. De esta manera se puede ver cuál es la probabilidad de que ocurra un evento con el cambio de dos situaciones, sea por una variable continua o *dummy*.

A diferencia de la regresión lineal que calcula los estimadores o coeficientes $\vec{\beta}$ mediante mínimos cuadrados ordinarios, la regresión logística no asume una relación lineal entre regresores y la respuesta y calcula dichos estimadores por Máxima-Verosimilitud (MV). Con MV maximizamos el logaritmo de verosimilitud mediante un procedimiento iterativo de optimización no lineal de Newton-Raphson.

Para las variables respuesta multiclase, recurrimos a la regresión logística ordinal. En este caso el modelo asume que el odd-ratio que establece el efecto de las variables explicativas es el mismo para cualquiera de las posibles comparaciones, tal que:

$$O(g) = \frac{P(Y \leq g|\vec{X})}{P(Y > g|\vec{X})} = e^{\beta_{Og}+\beta_1 X_1+\ldots+\beta_I X_I} \qquad \text{Ec.14}$$

Donde $\beta_{Og} \neq cte$ debido a que depende de g pero $\beta_1, \beta_2, \ldots, \beta_I = cte$ debido a que no dependen de g.

5.4.2.2 Principal Components Ordinal Logistic Regression with Lasso Penalization

Este procedimiento no es un modelo en sí mismo, es la unión resultante de convertir la matriz de datos **X** en estructuras latentes mediante **PCA** y aplicarle la regresión logística ordinal como hemos visto anteriormente (Kemalbay y Korkmazoğlu, 2014). Al convertir las variables originales en estructuras latentes eliminamos la colinealidad de las variables y condensamos la máxima variabilidad de los datos en **X** en las primeras componentes principales. No obstante, este método no sería adecuado si dicha variabilidad no se correlaciona con la respuesta. Los regresores que se emplean en los diferentes modelos que requieren una previa transformación de las variables originales en variables latentes emplean los *scores* $\mathbf{t_a}$.

5.4.2.3 Árboles de decisión

Los árboles de clasificación y regresión (en inglés *Classification and Regression Trees* o CARTS) son una técnica estadística basada en múltiples dicotomías o bifurcaciones anidadas en forma de árbol de manera que al seguir cada una de las ramas del árbol se obtenga una predicción para la clase de pertenencia (clasificación) o el valor que toman (regresión) los individuos a partir de las propiedades que se han decidido en las diversas bifurcaciones (Gey y Nedelec, 2005).

El árbol comienza con un único nodo en el primer nivel llamado nodo raíz. Los nodos situados al final del árbol se llaman nodos terminales. Todos los demás nodos en los niveles intermedios son nodos internos. Los nodos raíz e internos se bifurcan en dos nodos en el siguiente nivel (nodos hijos). El algoritmo construye un árbol de múltiples caminos y para cada nodo se busca la variable que aporte una mayor ganancia de información para la clase o el valor de salida. La forma en la que se clasifican los datos es por medio de una o varias reglas asociadas que satisfacen ciertas características y le dé da un sentido desde el punto de vista de las predicciones.

Una ventaja de los árboles de decisión en relación con las técnicas tradicionales es su gestión en casos donde existen estructuras discriminantes muy lejos de la linealidad.

5.4.2.4 Random Forest

Random Forest (Breiman, 2004) es una técnica basada en Árboles de decisión (algoritmo CART). La principal diferencia es basa en introducir en cada nodo un índice de aleatoriedad *p* de todas las variables y de éstas selecciona la mejor partición. Dicha partición sucede seleccionando individuos al azar mediante un muestreo con reemplazo, creando así diferentes sets de datos. A continuación, se crean árboles seleccionando variables al azar en cada nodo del árbol. De esta manera se crean diferentes árboles con cada set de datos que se había aleatorizado previamente y se predicen utilizando las nuevas observaciones (no empleadas) por cada árbol utilizando un voto mayoritario para el número de árboles que hagan una buena predicción de las nuevas observaciones antes mencionadas.

5.4.2.5 Support Vector Machine

El método de Support Vector Machine (Cortes y Vapnik, 1995) está basado en la minimización del riesgo estructural (SRM en inglés). Esta teoría permite escoger un clasificador que minimiza una cota superior sobre el riesgo y proporciona una buena medida que generalizan bien sobre otros datos. Esto significa que el modelo tiene una tendencia a evitar el sobreajuste respecto a otros métodos como redes neuronales.

Adicionalmente, se introduce una función Kernel que construye regiones de decisión no lineales para discriminar entre clases, dependiendo del Kernel escogido. Los datos con más información se llaman vectores soporte, y no inducen a mínimos locales.

El método muestra un alto rendimiento en diversas aplicaciones, no obstante, no es un método que ofrece información sobre las variables, su importancia, su relación con la variable respuesta, etc.

El modelo se basa en la hipótesis de que cualquiera de sus elementos se puede representar de manera numérica cuyo proceso puede representarse como una función de mapeo entre el conjunto original y el grupo de clases en un espacio de características de dimensión superior. SVM busca un hiperplano de separación que maximice el margen *m* entre las clases del espacio (Ilustración 7). Dicha maximización es un problema de programación cuadrática que puede ser resuelto mediante los multiplicadores de Lagrange. De esta manera, SVM busca el hiperplano óptimo sin ningún conocimiento sobre el mapeo usando los kernels.

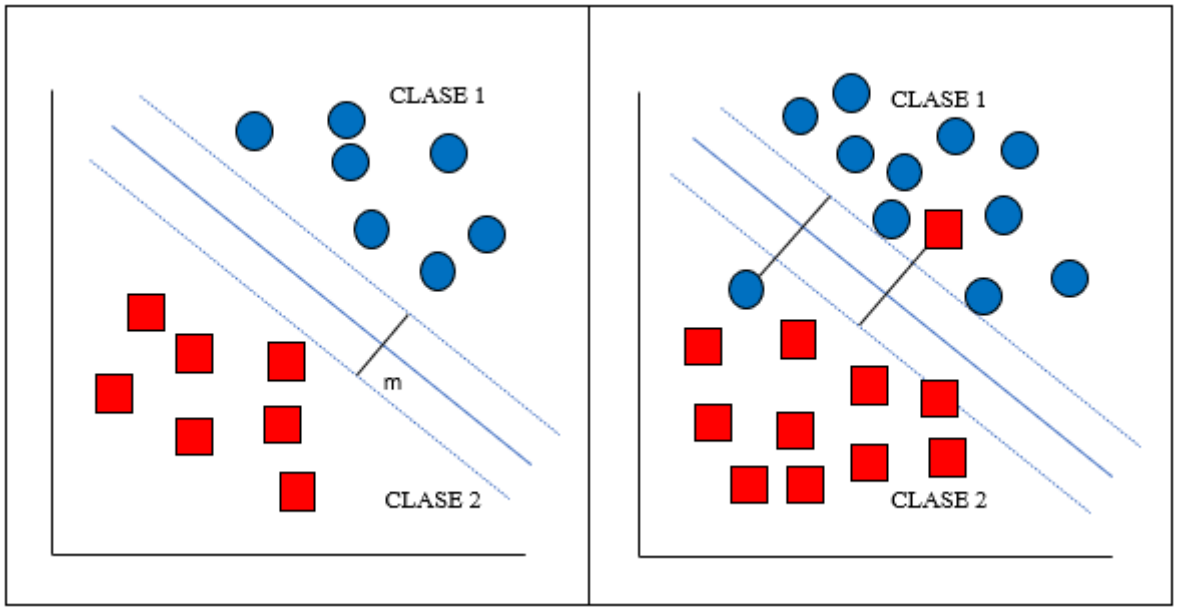


*Ilustración 7 Dos ejemplos separados por un Kernel lineal. A la izquierda, dos clases linealmente separables con infinitas soluciones a lo largo del vector **m**, con una única solución óptima (maximizando la separación entre las proyecciones). A la derecha, dos clases cuasi-separables linealmente, cuyo el valor óptimo sería el "maximal margin hyperplane" con gran riesgo de sobreajuste si no se parametriza correctamente el modelo. Fuente: elaboración propia.*

Para los casos linealmente separables, existe un número infinito de vectores **w** (no confundir con los vectores *loadings* de las direcciones principales empleado en otros modelos), pero solo existe un hiperplano óptimo, que maximiza el margen entre las proyecciones de los puntos de entrenamiento.

Para los casos que no son linealmente separables, se deben permitir "violaciones" a la clasificación en la formulación del modelo, tal que:

$$y_i(w \cdot z_i + b) \geq 1 - \xi_i, \text{ donde i=1, 2, ..., I} \qquad \text{Ec.15}$$

Para ello existe en la ecuación general de SVM el término $\sum \xi_i$ que se considera un error de clasificación, y debe tomar valores $\xi_i \geq 0$, teniendo en cuenta que cuando x_i no satisface la ecuación entonces $\xi_i \neq 0$. Finalmente, el problema de optimización vendría sujeto a lo expuesto anteriormente con la siguiente función objetivo:

$$min\left\{\frac{1}{2}w \cdot w + C\sum_{i=1}^{I}\xi_i\right\} \qquad \text{Ec. 16}$$

donde C es una constante que puede ser definida como un parámetro de regularización, también conocida como Coste y es el único parámetro libre de ser ajustado en la formulación del modelo SVM. Se puede definir como un balance entre la maximización del margen y la "violación" a la clasificación.

Habiendo comentado que C es el único parámetro libre de ser ajustado, existe otro parámetro muy común que también ha de ser ajustado, pero este no se encuentra en el modelo general de SVM sino en la función Kernel, concretamente en el Kernel Gaussiano. En la literatura se le llama en inglés *Gaussian Kernel* o *Radial basis function*

kernel (RBF). El parámetro del kernel es γ y define la distancia de las observaciones que se deja que influyan en el modelo.

Continuando con el modelo SVM, el hiperplano óptimo se considera un problema de programación cuadrática, que puede ser resuelto a partir de lo que se conoce como un lagrangiano donde aparecen multiplicadores de Lagrange positivos α_i y a partir de la siguiente expresión:

$$\bar{w} = \sum_{i=1}^{I} \bar{\alpha}_i y_i z_i \qquad \text{Ec.17}$$

Cumpliríamos los requisitos para calcular dicho hiperplano óptimo, como veremos a continuación:

$$\bar{w} \cdot z + \bar{b} \qquad \text{Ec.18}$$

Donde el escalar b puede determinarse de las condiciones del teorema de Kuhn-Tucker.

5.4.2.6 k-nearest neighbor

K-Nearest Neighbor es un método de clasificación no parámetrico (Peterson, 2009). De esta manera no se requiere hacer suposiciones de normalidad ni otras distribuciones estadísticas en los datos. El método opera tomando como base una medida de distancia entre puntos, siendo la distancia euclídea la más habitual. Cuando se introduce una observación que queremos clasificar, se calculan las distancias del punto que queremos clasificar respecto a todos los puntos de la muestra y se seleccionan los puntos más próximos a la clasificación que queremos realizar. Se calcula la proporción de los *k* puntos que pertenece a cada una de las poblaciones y se clasifica la observación en la población con mayor frecuencia de puntos en los *k* vecinos más cercanos.

K es un parámetro que ha de ser seleccionado manualmente. Normalmente se seleccionan varios y se escoge el que menor error de clasificación induce. De manera intuitiva, si se escoge un valor *k* elevado, cabe la posibilidad de condicionar la clasificación a la mayoría de las observaciones y no a la similitud de las más cercanas. De la misma manera, seleccionar un valor *k* pequeño puede inducir un sesgo si en los datos no hay una clara separación de clases y utilizamos como referencia otra clase.

5.4.3 Selección de variables

5.4.3.1 Eliminación recursiva de variables

La eliminación recursiva de variables es una estrategia altamente eficaz que consigue evitar la búsqueda exhaustiva de todas las combinaciones posibles de variables debido a que ello conlleva un gasto computacional muy elevado.

Dicho algoritmo selecciona un modelo, que en este caso ha sido el de *Random Forest*, que evaluará cada conjunto de predictores (Darst, Malecki, y Engelman, 2018). La condición requerida en el modelo utilizado es la posibilidad de ordenador las variables de mayor de menor importancia. En *Random Forest* se evaluaría por tanto la reducción del Error cuadrático medio (*Mean Square Error* en inglés, o MSE) de cada predictor. En un modelo de regresión lineal el criterio sería el estadístico *t* de los predictores, por ejemplo.

El modelo se ajusta con el conjunto de entrenamiento incorporando todas las variables disponibles. En el algoritmo se debe seleccionar el conjunto de variables. En este TFM todas las pruebas han usado todas las posibles combinaciones desde un mínimo de dos hasta un máximo de *j* variables. En cada conjunto de prueba se calcula el error y se obtiene un ranking de importancia de los predictores. Se reajusta el modelo con nuevos tamaños del modelo calculando el

error del nuevo modelo. Finalmente se selecciona el grupo de predictores de tamaño *n* que haya conseguido un error menor (Rodrigo, 2018).

Este proceso de selección presenta como ventajas simplificar el número de variables y eliminar aquellas que puedan no estar relacionadas con la variable respuesta. Como desventajas, en este proceso, la selección de predictores forma parte del ajuste del modelo, aumentando el riesgo de sobreajuste de este. Para solucionar esto, se puede incluir el algoritmo anterior en un bucle de validación cruzada que solo emplee los datos de entrenamiento. El error final de cada modelo se obtiene a partir de los errores obtenidos por cada conjunto de validación.

5.4.3.2 Penalización lasso

Lasso (least absolute shrinkage and selection operator) es un método de selección de variables y regularización que se para mejorar la exactitud, así como la interpretabilidad del modelo producido (Tibshirani, 2011). Originalmente se creó para el método de mínimos cuadrados (Regresión lineal), hoy en día Lasso está adaptado para una gran variedad de modelos.

La penalización *lasso* aplica una penalización absoluta a los coeficientes, también conocida como penalización ℓ1, forzando los coeficientes a tomar un valor de cero. Existen otros tipos de penalización como *ridge* que penaliza la suma de los coeficientes al cuadrado reduciendo su tamaño, conocida como penalización ℓ2. El hecho de que lasso fuerce algunos coeficientes a cero permite eliminar esas variables del modelo, simplificándolo. Como desventaja, lasso funciona mal si no existen variables relacionadas con la variable respuesta.

La función de pérdida para un modelo de regresión se definiría como:

$$\ell_1(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\sum_{j=1}^{m}\left|\hat{\beta}_j\right| \qquad \text{Ec.19}$$

Donde λ es un parámetro que controla la penalización, en el que si λ=0 no se aplica ningún tipo de penalización, y si λ= ∞ todos los coeficientes son 0. Por tanto, λ es un parámetro que se optimizar. Mediante la validación cruzada se puede seleccionar un valor de lambda con menor error.

5.4.3.3 VSURF: Variable Selection Using Random Forests

Esta metodología de selección de variables se centra en la potencia del algoritmo de *Random Forest* en el uso estratégico del error OOB (out of bag o *bagging* en inglés) y la importancia de las variables (Genuer, Poggi, y Tuleau-Malot, 2015).

El *bagging* hace referencia al muestreo repetido (*bootstrapping*) para reducir la varianza del algoritmo CART. Esto es efectivo debido a que la varianza de la media de un grupo de observaciones es σ^2/n, donde n es el número de muestras, y promediando un conjunto de observaciones se reduce la varianza y aumenta la precisión.

Por otro lado, al generar multiples árboles con distintos tamaños de muestra y variables se puede recurrir a otras estrategias que cuantifiquen la importancia de las variables: el incremento del MSE (*Mean Square Error* para regresión) o MER (*Misclassification Error Rate* para clasificación) y el incremento de la pureza de nodos.

En primer lugar, en el incremento del MSE/MER se puede observar la diferencia del error del modelo sin una variable respecto al error del modelo estándar, tal y como se muestra a continuación:

$$VI(X^j) = \frac{1}{ntree}\sum_t\left(\widetilde{errOOB_t^j} - \widetilde{errOOB_t}\right) \qquad \text{Ec.20}$$

Donde VI es la importancia obtenido a partir de la diferencia del error de regresión o de clasificación obtenido tras la permutación de una variable ($\widetilde{errOOB_t^j}$) y el error inherente del modelo de un solo árbol *t* en una muestra $\widetilde{errOOB_t}$.

En segundo lugar, el incremento de la pureza de nodos cuantifica el incremento total en la pureza de los nodos a causa de las divisiones en las que aparece el predictor (como promedio de todos los árboles). Se calcula registrando el descenso conseguido en la medida empleada como criterio (índice Gini, entropía o *MSE/MER)* en cada división de los árboles.

El algoritmo VSURF emplea exclusivamente las medidas de reducción del MSE/MER.

El algoritmo de selección se divide en dos objetivos de selección, uno de *thresholding* y otro de predicción. No obstante, se utiliza el mismo criterio de selección para ambos. En el *thresholding* se filtran las variables que están altamente relacionadas con la variable respuesta incluso si existe algo de redundancia (colinealidad). En la predicción se trata de reducir la lista a un pequeño número de variables con muy poca redundancia que produzcan una buena medida de predicción de la variable respuesta. La diferencia entre *thresholding* y predicción en este caso no es explícitamente el significado literal de dicha terminología, sino el grado de parsimonia. Unas veces nos puede interesar identificar el mayor número de variables y otras simplificar un modelo, según los objetivos particulares de cada investigación.

En el segundo paso de selección de variables, el de predicción, encontramos una subdivisión en la que ofrece dos resultados, con diferente grado de parsimonia. Se puede ver como una fase de selección basado en la predicción en el que tras obtener un primer resultado se trata de depurar todavía más.

5.4.4 Optimización de hiperparámetros

Para las técnicas estadísticas de aprendizaje supervisado, normalmente se usan modelos paramétricos donde p(y|X,Θ) para explicar nuestros datos e inferimos valores óptimos del parámetro Θ como por ejemplo por máxima verosimilitud (*Maximum* Likelihood o ML en inglés) o mínimos cuadrados ordinarios (*Ordinary least squares* o OLS en inglés). Cuando aumenta la complejidad de nuestros datos a veces se requiere utilizar métodos que expliquen mejor estos datos. Con frecuencia estos modelos se construyen con un mayor número de parámetros para realizar un mejor ajuste de estos datos. Si los métodos utilizan un número fijo de parámetros se llaman métodos paramétricos.

Por otro lado, los métodos no paramétricos utilizan un número de parámetros que crecen junto con el tamaño de los datos. Por ejemplo, KNN es un método estadístico no paramétrico porque el modelo crece con el tamaño de los datos de manera que cada punto de entrenamiento es un parámetro en el modelo y esto influirá en las predicciones, además, K es un hiperparámetro porque se debe ajustar de manera libre.

En este proyecto se han utilizado dos metodologías para ajustar hiperparámetros de los modelos. Por la facilidad de implementación en R se han usado

5.4.4.1 Proceso Gaussiano

El proceso gaussiano (GP) es una técnica basada en Procesos estocásticos gaussianos e Inferencia bayesiana (Wu et al., 2019). Es además una generalización de la Distribución de probabilidad de Gauss. Mientras una distribución de probabilidad describe variables al azar para distribuciones multivariantes, existe un proceso estocástico que controla las propiedades de las funciones. Un proceso gaussiano es un proceso estocástico que supone que una colección finita de variables posee una distribución normal multivariante. De la misma manera que la distribución normal, GP se define por su media $m\ x\ \rightarrow \mathbb{R}$ y su covarianza k: $x \times\ x\ \rightarrow \mathbb{R}$, de manera que:

$$f(x) \sim GP(m(x), k(x, x')) \qquad \text{Ec. 21}$$

En la que la función de densidad *f(x)* para un valor aleatorio de *x* no es un escalar sino una función de distribución normal sobre todos los posibles valores de *f(x)*. Por conveniencia se asume que la media de un PG es m(x) = 0. Para la función de covarianza *k*, se suele utilizar la función cuadrática exponencial (RBF):

$$k(x_i, x_j) = exp\left(-\frac{1}{2}\left\|x_i - x_j\right\|^2\right) \qquad \text{Ec. 22}$$

donde x_i y x_j representan las observaciones i-ésimas y j-ésimas, respectivamente. Cuando los valores son parecidos la función *k* se aproxima a 1, si se alejan mucho, se aproximará a 0. Intuitivamente podemos decir que si los valores son cercanos es porque tienen una fuerte correlación.

La posterior distribución se obtiene a partir del set de entrenamiento suponiendo una Distribución normal multivariante donde $f \sim (0, \boldsymbol{K})$, donde $\boldsymbol{K}$ se calcula con la ecuación 22. Según la función f se calcula el valor de la función con una nueva observación x_{t+1}, donde $f_{t+1} = f(x_{t+1})$. En los procesos gaussianos, $f_{1:t}$ junto con f_{t+1} siguen una distribución normal dimensional de *t*+1.

Si el set de entrenamiento es lo suficientemente grande, GP puede obtener una estimación de la distribución de la función $f(x)$ aproximada. Para el caso de la optimización de hiperparámetros, ***y*** o $f(x)$ sería la medida del error obtenido por validación cruzada en base a la combinación de los distintos hiperparámetros.

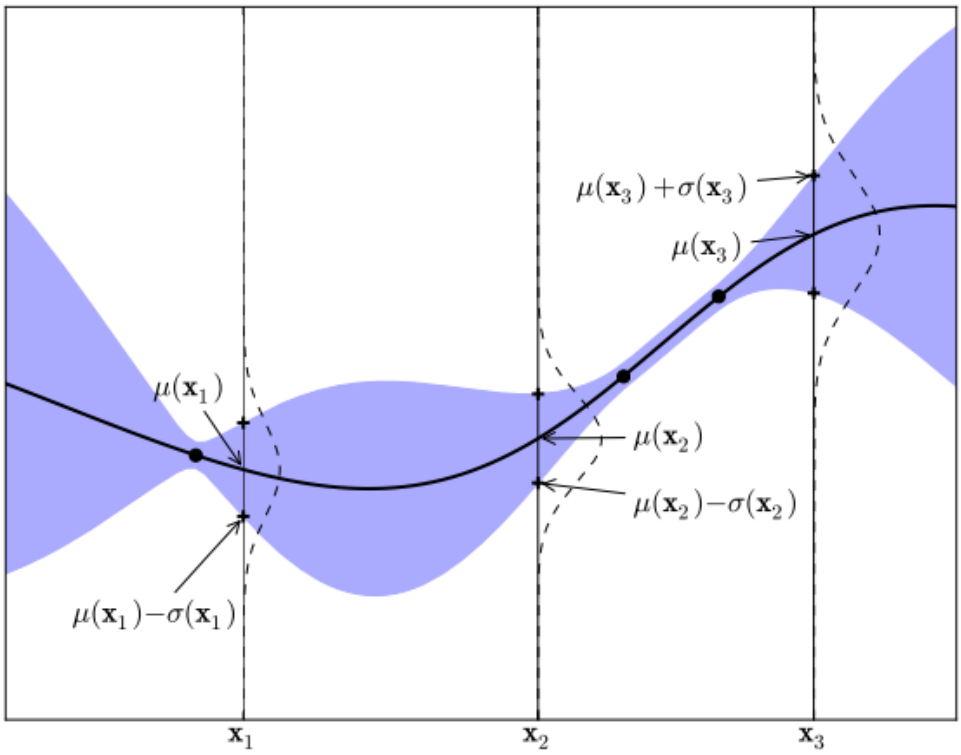


Ilustración 8 Proceso Gaussiano simple de una dimensión con tres observaciones. La línea negra oscura es la media de GP en la predicción ofrecida por la función objetivo con unos datos determinados y el área azul muestra la varianza alrededor de dicha media. Las líneas verticales corresponden a la media y la desviación estándar de GP en los puntos de predicción $\boldsymbol{x_{1:3}}$. ***Fuente:*** *(Brochu, Vlad, y De Freitas, 2010)*

Tras la obtención de la distribución de la función objetivo, se puede utilizar la optimización Bayesiana para maximizar la función de adquisición, que determina dónde se espera que sea de mayor utilidad llevar a cabo la siguiente la siguiente observación para minimizar el número de pasos que deben llevarse a cabo. Existen varias funciones que se pueden utilizar como función de adquisición, las más comunes son la de mejora de la probabilidad, de mejora esperada y de límite superior de confianza (Wilson y Adams, 2013). Para el presente trabajo se hará uso de la última.

5.4.4.2 Grid Search

Esta técnica realiza de manera sistemática múltiples pruebas con el modelo de aprendizaje supervisado en cuestión a partir de un rango de valores de cada hiperparámetro definido por el usuario (Lerman, 1980). La selección de los hiperparámetros óptimos vendrá determinada por la combinación de ellos que haya obtenido una precisión más alta en el modelo. La ventaja principal es que mapea el espacio propuesto por el usuario y según el tamaño del rango y los intervalos ofrece un gran potencial en la obtención del óptimo. Como contrapartida puede ser muy costoso computacionalmente para un gran número de hiperparámetros y rangos muy grandes.

5.4.5 Evaluación de resultados

5.4.5.1 Evaluación mediante validación cruzada

Para la evaluación de los modelos estadísticos se ha empleado el método de Validación cruzada dejando uno fuera, más conocido en inglés como *leave one out* (LOO o LOO-CV).

La técnica LOO-CV es un método iterativo que utiliza todas las observaciones excepto una que se utiliza como validación (Kulesa, Krzywinski, Blainey, y Altman, 2015). El proceso continúa hasta que se realiza tantas veces como observaciones existen dejando fuera del modelo una observación de manera consecutiva.

Esto permite reducir la variabilidad que se origina respecto a otras técnicas como *Hold out* o método de retención, que divide los datos disponibles en dos grupos: sets de entrenamiento y validación. Esto es debido a que en un *Hold out* podríamos estar eligiendo al azar muestras de distintas clases que entrenarían el modelo de manera distinta para una partición determinada. Al evitar esta separación aleatoria, los resultados de LOO-CV son completamente reproducibles.

Es un proceso muy costoso computacionalmente, pero para bases de datos relativamente pequeñas como la que ocupa en este trabajo es una técnica adecuada para poder entrenar modelos con el máximo número de observaciones.

5.4.5.2 Análisis de la Varianza (ANOVA)

El análisis de la varianza permite observar diferencias estadísticas significativas entre los métodos de regresión para comparar el rendimiento del modelo. También se ha utilizado para observar diferencias entre los niveles de los distintos sustratos patológicos y la variable continua con la que se relaciona (Ej.: Fibrosis vs CPA, Inflamación vs CD45, etc.).

La técnica estudia la significación de uno o más factores comparando las medias de los niveles de los factores (Girden, 1992). Para poder realizar esta técnica se deben cumplir tres hipótesis: normalidad de los datos, independencia de los residuos, homocedasticidad de los residuos.

5.4.5.3 Prueba de Kruskal-Wallis

En esta prueba se persigue el mismo objetivo que en el ANOVA: observar diferencias estadísticas entre las distintas poblaciones, a excepción de que es un método no paramétrico y la única suposición es que los datos deben provenir de la misma distribución, es decir, debe haber homocedasticidad (McKight y Najab, 2010). La hipótesis nula sostiene que los distintos grupos pertenecen a la misma población, y la hipótesis alternativa (con un p-valor < 0.05) indicaría que los grupos pertenecen a distintas poblaciones.

5.4.5.4 Prueba de los rangos con signo de Wilcoxon

Esta prueba estadística permite comparar poblaciones cuando sus distribuciones no se ajusten bien a otras pruebas paramétricas. Es una alternativa al *t-test* de muestras dependientes cuando las muestras no siguen una distribución normal, bien porque realmente no la tienen o bien porque la muestra es tan reducida que no se puede determinar si realmente proceden de poblaciones normales.

La prueba de los rangos con signo de Wilcoxon compara si las diferencias entre pares de datos pertenecen a la misma población, de manera que mide las diferencias entre cada par de observaciones, trabajando sobre rangos de orden, por lo que solo se puede usar a variables ordinales. Como punto a favor, ignora los valores extremos, (los *t-test* trabajan con medias y tienen en cuenta valores extremos), ya que utiliza las medianas (Woolson, 2007).

5.4.5.5 Variables respuesta binarias: Curvas ROC y PR

En los sistemas de clasificación binarios es habitual utilizar las curvas ROC y PR para observar el rendimiento y la eficacia de un modelo según distintos parámetros. Para ello extraemos del modelo la probabilidad de pertenencia a una clase. A estos modelos se les llama clasificadores probabilísticos y con las probabilidades podemos crear una curva variando el umbral de valores. Para calcular el umbral óptimo se empleará el índice de Youden, como se verá más adelante.

Para explicar en qué consisten las curvas ROC y PR antes hay que introducir los términos que encontraríamos en una matriz de confusión (Ilustración 9): Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN).

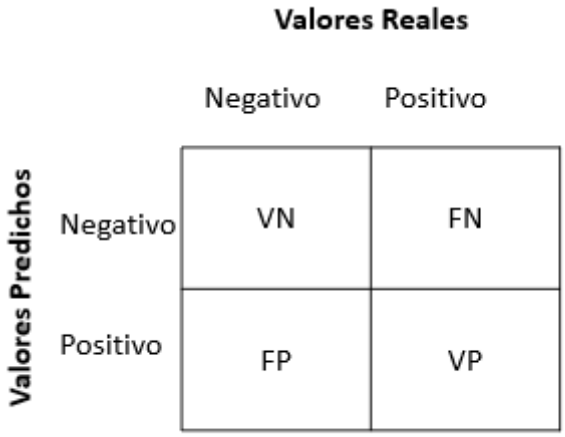


Ilustración 9 Matriz de confusión tipo con formato 2x2. En las filas se ordenan los valores predichos por el algoritmo entrenado y en las columnas los valores reales. Por tanto, los aciertos se organizan en la diagonal de la matriz. Fuente: Elaboración propia.

5.4.5.5.1 Curva característica operativa del receptor

La Característica Operativa del Receptor (COR o ROC en inglés) representa gráficamente la sensibilidad frente a la especificad según se varía el umbral de discriminación (Fan, Upadhye, y Worster, 2006). La sensibilidad es la proporción de observaciones predichas como positivas de todas las verdaderas positivas $VP/(VP + FN)$ y de manera similar la especificidad es la proporción de observaciones predichas incorrectamente como positivos de las observaciones que son negativas $1 - (FP/(VN + FP))$.

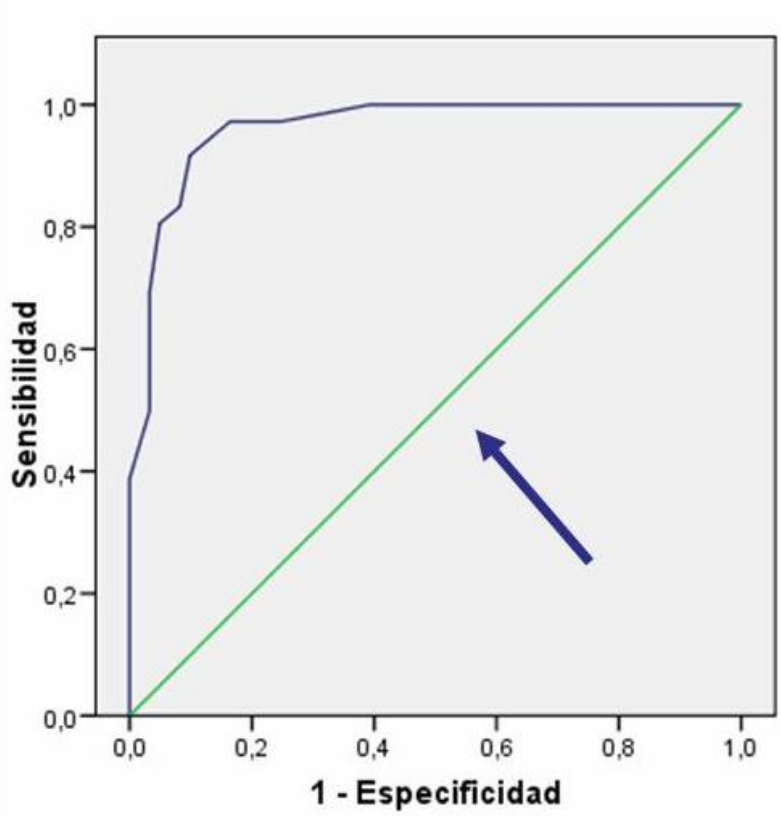


Ilustración 10 Curva ROC característica de una buena clasificación con una sensibilidad y especificidad, muy por encima de la diagonal que señala la flecha morada. Fuente: (App4Stats, s.f.)

La curva ROC muestra la relación entre sensibilidad y especificad (Ilustración 10). De manera intuitiva, se puede decir que la curva ROC muestra la relación entre lo bueno que es el modelo detectando VP frente a los VN. Un resultado es mejor cuanto más se acerque la curva a la esquina superior izquierda y normalmente se muestra un umbral en forma de diagonal que representa que nuestro modelo es muy mal clasificador.

Para generalizar los resultados, normalmente se recurre al área bajo la curva (AUC en inglés), de manera que podemos generalizar el rendimiento de un modelo clasificador, no obstante, en la práctica en ocasiones existen regiones específicas donde se obtiene un peor rendimiento.

En situaciones con clases desbalanceadas la curva ROC tiende a mostrar resultados demasiado optimistas. Es decir, si tenemos muchos controles (clase 0) y pocos casos (clase 1), es posible que el clasificador tienda por defecto a apuntar hacia la clase 0, puesto que tendremos un sobreajuste para esta categoría, de esa manera tendremos muchos VN y valores muy bajos del resto, forzando una sensibilidad "falsamente" elevada. Para estos casos es bueno recurrir a la curva PR, explicada a continuación.

5.4.5.5.2 Curva Precision-Recall

La curva *precision recall* (PR) representa gráficamente la precisión frente al *recall* o sensibilidad (Flach y Kull, 2015). La precisión es la proporción de clasificaciones correctas de todas las que nuestro clasificador considera como correctas $VP/(VP + FP)$. La sensibilidad ya fue definida anteriormente como la proporción de verdaderos positivos de todos los positivos. De esta manera las métricas están relacionadas de modo que si se consigue aumentar la precisión disminuirá el *recall* y viceversa.

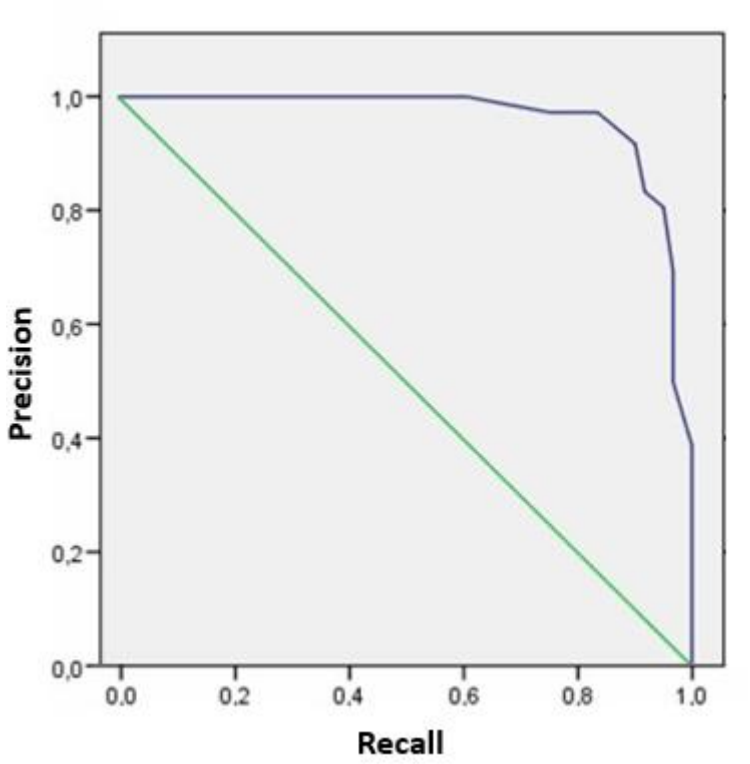


Ilustración 11 Curva de Precision-Recall. La interpretación es contraria a la curva ROC, cuanto más arriba a la derecha, más buena es la predicción. En este caso representamos la precisión frente al recall (o sensibilidad). Fuente: Elaboración propia.

En este caso los mejores resultados se aproximan a la esquina superior derecha (Ilustración 11). Existe también una diagonal de referencia que indica resultados pobres en los que la precisión equivale al *recall*.

Es posible generalizar el rendimiento del clasificador a partir de la PR AUC o área bajo la curva PR. Cuanto más cercano a 1 sea el valor, mejor será el modelo en cuestión.

5.4.5.5.3 Índice de Youden

El índice de Youden es una técnica estadística que determina el rendimiento de una prueba de diagnóstico que sea dicotómico (Fluss, Faraggi, y Reiser, 2005). Se utiliza normalmente de manera conjunta con la curva ROC. Se puede definir como:

$$J = Especifidad + Sensibilidad - 1 \qquad \text{Ec. 23}$$

Puede tomar valores de -1 a +1 y cuando la prueba da 0 da a entender que la prueba de diagnóstico ofrece exactamente la misma proporción de resultados positivos para los grupos control y casos. Un valor de 1 indica que no hay falsos positivos o falsos negativos. Es equivalente al área bajo la curva en un único punto operativo.

5.4.5.6 Variables respuesta no binarias: Método de Obuchowski

Para las variables respuesta categóricas con más de dos clases o multiclase, se han invertido diversos esfuerzos por adaptar la curva ROC a estos sistemas. Existen dos opciones ampliamente utilizadas para los sistemas multiclase: "una contra una" y "una contra todas". En la primera, cada clase se compara por pares frente a las demás, manteniendo las propiedades habituales de una curva ROC con la principal desventaja de generar tantas curvas como combinaciones entre las distintas clases existan, a razón de $n^2 - n$. Para la segunda, cuando comparamos una frente al resto tenemos

la ventaja de la complejidad aumenta de forma lineal, por cada clase n de más solo se requiere otra curva ROC frente al resto, no obstante, estamos sesgando la información al estar comparando con diversas clases a la vez (Wandishin y Mullen, 2009).

Existen diversas situaciones en las que se puede utilizar una variable indicadora como *gold standard* de diversas patologías. Por ejemplo, el *gold standard* para dolor abdominal agudo se puede utilizar como indicador para la apendicitis, gastroenteritis, constipado, obstrucción intestinal e infección del tracto urinario (Obuchowski, Goske, y Applegate, 2001). El método de actuación consiste normalmente en dicotomizar los resultados de esta escala (una clase frente a las demás) y aplicar el método ROC tradicional. No obstante, esto conlleva una estimación sesgada de la precisión (Obuchowski, 2005).

En este Trabajo de Fin de Máster se propone el uso del método de Obuchowski para problemas multiclase con el fin de no recurrir a estrategias que nos puedan hacer perder información o dificulten la interpretación en los análisis. Este método de evaluación puede ser especialmente útil para el campo de la medicina basada en niveles de una enfermedad.

El método de Obuchowski ofrece una estimación no paramétrica basado en la función lineal de Kendall's τ (Hanley y McNeil, 1982) y una extensión de la estimación de la precisión de Wilcoxon-Mann_Whitney (Snedecor y Cochran, 1967), haciendo posible el cálculo de la precisión para escalas en las que el *gold standard* es continuo, ordinal o nominal.

Donde los diferentes estimadores $\hat{\theta}$ se interpretan de manera ligeramente distinta según el tipo de escala utilizada (continua, ordinal y nominal):

- En variables binarias $\hat{\theta}_{bin.}$ se interpreta como la probabilidad de puntuar con un valor más alto a un paciente que es un verdadero positivo respecto a un verdadero negativo para dos pacientes elegidos al azar de cada clase.
- En variables continuas $\hat{\theta}_{cont.}$ es la estimación de la probabilidad de que el resultado de la prueba diagnóstica coincidirá con el verdadero orden de los pacientes.
- En variables ordinales $\hat{\theta}_{ord.}$ es la estimación de la probabilidad de que, si se elige a un paciente elegido al azar de cada clase, los pacientes clasificados de una clase mayor estarán realmente diagnosticados a una clase superior más que un paciente que pertenezca a una clase menor.
- En variables ordinales $\hat{\theta}_{nom.}$ es la estimación de la probabilidad de que los pacientes pertenezcan correctamente a la categoría predicha por el modelo.

Tabla 1 Estimadores para el test diagnóstico de precisión.

Tipo de escala	Estimador de la precisión	Condiciones	Terminología
Binaria	$\hat{\theta}_{bin.} = \frac{1}{n_t n_s} \sum_{i=1}^{n_t} \sum_{j=1}^{n_s} \psi(X_{it}, X_{js})$	$\psi = 1$ si $X_{it} > X_{js}$ $\psi = 0.5$ si $X_{it} = X_{js}$ $\psi = 0$ si $X_{it} < X_{js}$	t = verdaderos positivos s = verdaderos negativos X_{it} = paciente i-ésimo de t X_{js} = paciente j-ésimo de s
Continua	$\hat{\theta}_{cont.} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \psi(X_{it}, X_{js})$	$\psi = 1$ si $t > s$ y $X_{it} > X_{js}$ $\psi = 1$ si $s > t$ y $X_{jt} > X_{is}$ $\psi = 0.5$ si $t = s$ o $X_{it} = X_{js}$ $\psi = 0$ de cualquier otra manera	N = número total de pacientes o muestras
Ordinal	$\hat{\theta}_{ord.} = 1 - \sum_{t=1}^{T} \sum_{s>t}^{T} w_{ts} L(t,s)(1 - \hat{\theta}_{ts.})$	$\hat{\theta}_{ts.}$ Es el estimador de la escala binaria por pares	T = número total de estados de enfermedad L = función de penalización w = función de pesos de los estados
Nominal	Equivalente al ordinal	$\hat{\theta}_{ts} = \frac{1}{n_t n_s} \sum_{i=1}^{n_t} \sum_{j=1}^{n_s} \psi(D_{(t-s)tj}, D_{(t-s)sk})$ $\psi(D_{(t-s)tj}, D_{(t-s)sk}) = 1$ si $D_{(t-s)tj} > D_{(t-s)sk})$ $\psi(D_{(t-s)tj}, D_{(t-s)sk}) = 0.5$ si $D_{(t-s)tj} = D_{(t-s)sk})$ $\psi(D_{(t-s)tj}, D_{(t-s)sk}) = 0$ si $D_{(t-s)tj} < D_{(t-s)sk})$	D=diferencia entre puntuaciones de confianza de los estados o clases

En la *Tabla 1* podemos observar la formulación expuesta para las diferentes estimaciones según la escala utilizada. Para las escalas nominal y ordinal, podemos ver el término w_{ts} , que se refieren al peso de los estados. El peso de los estados se define mediante la siguiente expresión:

$$w_{ts} = \frac{n_t n_s}{\sum_{t=1}^{T} \sum_{j>i}^{T} n_i n_j} \qquad \text{Ec.24}$$

En los que T es el número total de categorías o estados de una enfermedad (nominales u ordinales) en la escala definida por el *gold standard*, de la manera que $t = 1,2, \dots, T$. Los estados de enfermedad se definen por t y s para los correctamente clasificados y los incorrectamente clasificados, respectivamente, y el número de pacientes en cada caso sería n_t y n_s.

La función de penalización es un parámetro que se debe ajustar manualmente. En la librería utilizada, su valor por defecto su valor es 1. Los valores que toma van entre 0 y 1, en los que 0 es una penalización nula y 1 es una penalización máxima. Puede ser de interés utilizar valores intermedios si no se quiere penalizar demasiado la equivocación entre clases vecinas. En un artículo de Obuchowski (2005) se sugiere una penalización fraccionada basada en el número de clases. Esta recomendación es la seguida para este TFM, y las matrices se pueden consultar en los Anexos 1, 2 y 3.

5.5 Historial de análisis

En esta sección se expone la información necesaria para situar al lector en el contexto de los análisis realizados.

5.6 Aclaraciones previas

En primer lugar, el análisis de todos los sustratos patológicos ha dado lugar a una gran cantidad de resultados de análisis. Para no exceder el volumen de este Trabajo de Fin de Máster, se presentarán los resultados más relevantes. Tampoco se incluirán en el apartado de Anexos por el mismo motivo. No obstante, todos los documentos de programación en R y de resultados quedan a disposición de cualquier interesado[10].

Para ofrecer una descripción más detallada, la base de datos se ha actualizado varias veces debido a la limpieza de datos y corrección de errores, así como debido a la revisión de la clasificación de algunas observaciones, y de algunos valores de las variables de imagen médica, por lo que los análisis se han repetido iterativamente. Se han generado 117 archivos con relación a este TFM, de los cuales 106 son scripts en R. Cada script tiene una media aproximada de 300 líneas de código. Existen algunos scripts con 3000 líneas y otros con 100. En cualquier caso, esta información se expone para explicar la inviabilidad de considerar este contenido como parte de los Anexos. Los resultados obtenidos de los distintos modelos se han organizado en distintos archivos MS Word y la extensión oscila de las 40 páginas a las 500, por lo que tampoco se incluirá este material en los Anexos.

5.7 Fases de investigación

En el artículo en revisión de Martí-Aguado et al. (2020) se realizaron unos análisis con la información disponible de entonces. Dicha información era una preselección de variables de RM y algunas covariables médicas (las mismas que las que se expondrán a continuación como las utilizadas en la fase 2 del TFM). Se evaluó la Fibrosis mediante la regresión logística ordinal con variables originales y variables latentes y selección de variables mediante el procedimiento *stepwise*, con buenos resultados.

5.7.1 Fase 1

Con ello, se decidió probar por otras vías en el momento de inicio de este TFM, a partir de otras variables ya completadas. En esta fase las variables a considerar fueron las variables de RM completas, además de edad, sexo, índice de masa corporal (IMC), APRI, FIB4, BARD, NAFLD, HSI, BAAT, ALBI, forns, TE Kpa, TE CAP. Estas variables se escogieron por su posible relevancia según el criterio técnico-médico. No existían muchos casos completos (aproximadamente 49 casos completos, según la variable respuesta del sustrato patológico a analizar), por lo que se decidió realizar un doble análisis de los casos completos con dichas covariables (~49 observaciones para las 93 variables) y sin ellas (~94 observaciones para las 80 variables de RM).

Las variables respuesta de los sustratos patológicos a considerar fueron las siguientes:

- Continuas: *Collagen Proportional Area* (CPA), *Cluster of differentiation 45* (CD45), Iron (Hierro), *Fat Proportional Area* (FPA).
- Categóricas:
    - Fibrosis: Score, Score dicotomizada, ISHAK e ISHAK dicotomizada.
    - Inflamación: Lobular, Lobular dicotomizada, Portal, Batts Ludwig, Batts Ludwig dicotomizada e Inflamación Avanzada.
    - Grasa: Esteatosis.
    - Hierro: Scheuer, Deugniers.

Los resultados de los análisis de esta fase no cumplieron con las expectativas generadas, por la calidad de los modelos obtenidos. Además, el rango de las matrices era bastante deficiente, impidiendo los análisis en determinadas

[10] E-mail de contacto: jorllum1@posgrado.upv.es

condiciones (a causa de la relación entre número de observaciones y de variables). Las técnicas multivariantes como *sPLS* y *sPLS-DA*, entre otras, se comportaron bien frente a la deficiencia de rango de la matriz de datos, a diferencia de otros métodos estadísticos como *Random Forest* y Regresión Logística (Binomial y Ordinal). Otro aspecto relevante fue la dificultad estadística de analizar variables muy desbalanceadas como ISHAK y Deugniers (Tabla 2), en las que al aplicar validación cruzada LOO existían clases que no se habían entrenado en el modelo. Adicionalmente, *Random Forest* genera un modelo a partir de la selección aleatoria de observaciones y variables, por lo que presentaba dificultades en la generación del modelo por el propio desbalanceo de la variable ISHAK aún con el modelo restringido.

Con todo esto, se decidió proceder a una segunda fase de análisis, reanudando los análisis tras los resultados del borrado de Martí-Aguado et al., (2020).

5.7.2 Fase 2

En esta segunda fase, se decidió centrar el análisis en las variables categóricas, descartando así las variables continuas por el momento, aunque se tendrán en cuenta para futuras líneas de trabajo. La justificación se debe a que trabajar con clases facilita la comprensión inmediata de la enfermedad, con una categoría basada en unos criterios. Trabajar con valores cuantitativos exige establecer puntos de corte en relación con los distintos grados de los sustratos hepatopatológicos, cuyos criterios[11] deben definirse con anterioridad.

En cuanto a las variables, se descartaron todas las covariables médicas a excepción de la edad, sexo e IMC, y se incluyeron síndrome metabólico, hipertensión arterial (HTA), diabetes (DM) y dislipemia (DL). El número de casos completos era de 93-97 observaciones, según el sustrato patológico analizado.

Se decidió descartar los sustratos de Fibrosis ISHAK, la inflamación avanzada y de hierro deugniers, por su baja eficacia y problemas estadísticos generados (pocas observaciones en algunas clases y/o clases muy desbalanceadas). Así mismo se decidió dicotomizar las variables de esteatosis y Scheuer.

En este proceso iterativo se cambiaron las variables para cada sustrato patológico, puesto que en la literatura se descubrió que algunas variables estaban altamente relacionadas con algún sustrato patológico en concreto, y no tanto para otros. Este es el caso de la esteatosis para las variables de resonancia magnética FF en su conjunto, y para hierro Scheuer R2W en su conjunto (França, 2017). De esta manera, con un número de variables menor, aumentaron los casos completos a 97 observaciones.

Se decidió así mismo eliminar estas variables para los sustratos patológicos restantes (Fibrosis e inflamación), por lo que estos sustratos se analizaron con 67 variables, 60 de RM y 7 covariables médicas.

6 Resultados

Antes de realizar ningún tipo de análisis se ajustó una semilla con valor 12345 para facilitar la reproducibilidad de esta investigación.

6.1 Limpieza de Datos

En la exploración visual inicial no se detectaron ningún tipo de anomalías a priori, no obstante, en R es muy útil la función *str()* para observar la clase a la que pertenece un objeto. Normalmente, cuando el programa no reconoce un patrón en los datos, suele considerar su clase como *character*. Se descubrió de esta manera que las variables APRI y FIB-4 aparecían con puntos y comas alternados como patrón decimal. En R el decimal se corresponde con el punto, no

[11] Los criterios actuales se basan en consideraciones morfo-anatomopatológicas que pueden producir distintos valores. Así, un paciente con inflamación Lobular de grado 3 podría obtener una puntuación de 0 en los criterios de Batts Ludwig, por eso elegir cuidadosamente los criterios es crucial en esta investigación.

obstante, tiene en cuenta que cuando se carga un archivo Excel, esto no es así. En cualquier caso, generaba un error, y con la exploración visual de estas variables se detectó el problema.

6.2 Análisis Exploratorio

Gracias al análisis exploratorio, se encontraron algunas anomalías que permitieron realizar correcciones de la base de datos.

La función *summary()* de R para variables continuas nos ofrece una serie de parámetros que nos permiten hacernos una idea de los datos a partir de la media, mediana, primer y tercer cuartil y valores mínimo y máximo, así como los valores faltantes. Realizando un summary de todas las variables, se observó que las variables cT1_std y T1_std tenían los mismos valores.

DS	DT	DU	DV	DW	DX
cT1_std	cT1_median	cT1_p2	cT1_p75	T1_mean	T1_std
72,8333211	886,909181	890,711452	901,23433	950,873338	72,8333211
34,1410111	739,927095	795,268763	705,102789	654,446772	34,1410111
82,5228052	796,496899	771,901224	822,971288	990,629649	82,5228052
58,6169476	648,202111	655,07284	656,579844	725,914301	58,6169476
55,5305688	874,280039	865,87584	878,301393	1014,24816	55,5305688
54,8906946	648,203573	669,25927	632,372064	634,643633	54,8906946
61,9913853	1173,83501	1191,63461	1171,08366	1155,85152	61,9913853
67,1390883	948,926055	961,175528	953,372326	1008,71992	67,1390883
76,6563426	809,145896	792,134397	829,65254	927,014265	76,6563426
67,1516043	1035,89158	1061,79754	1022,34706	991,647098	67,1516043
58,760036	720,581308	716,045079	734,813874	871,684512	58,760036
73,4559344	1307,36795	1322,11149	1317,18201	1293,88319	73,4559344
53,3785893	491,715377	486,81629	500,260171	669,636443	53,3785893
59,5882122	1145,17901	1190,26843	1120,55375	1059,8065	59,5882122

Ilustración 12 Las variables de RM cT1_std y T1_std tenían los mismos valores. El parecido en el nombre de las variables indujo posiblemente a error a la hora de pegar resultados.

De la misma manera, se observaron anomalías en otras de las variables de resonancia magnética, que presentaban unos máximos muy alejados de la media. Para algunas de las variables del grupo R2W, se observa una diferencia de escala en algunas observaciones que nos muestra los gráficos XY de la siguiente manera:

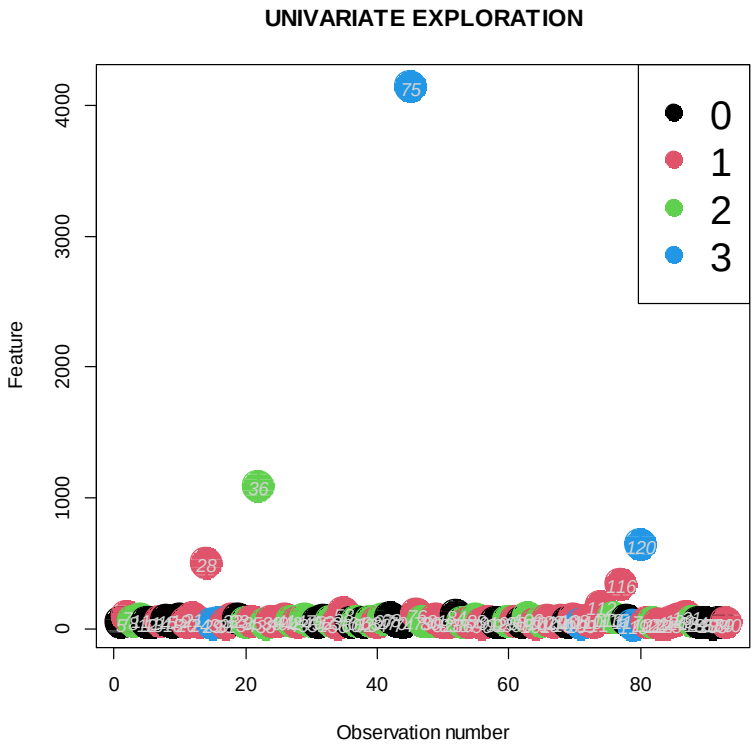


Ilustración 13 Gráfico XY de las observaciones o pacientes frente al valor de la variable R2W_mean. Los valores han sido previamente centrados y escalados, por lo que se observan más fácilmente las diferencias.

Para el caso de la R2W_std en la imagen 13, observamos que aparecen que las observaciones NT 120, 28, 116 y 112 con estimaciones muy elevadas respecto a la media de 0.

No se puede extraer una conclusión de dichas anomalías, puesto que se notificó al técnico encargado del análisis de imagen digital y en la comprobación de resultados aparecieron los mismos valores, por lo que se descarta cualquier incidencia en el traspaso de datos. Se requieren análisis más profundos para determinar este tipo de anomalías, relacionadas con la obtención de imágenes de resonancia magnética.

En cualquier caso, esto no afecta, en principio, a los análisis generados, puesto que en la fase 2, para la Fibrosis, inflamación y grasa se eliminaron estas variables, además, en las técnicas multivariantes se detectan este tipo de anomalías. Para el resto de los métodos sí que se ha procedido a la eliminación de estas observaciones, puesto que podrían generar un *leverage* muy elevado.

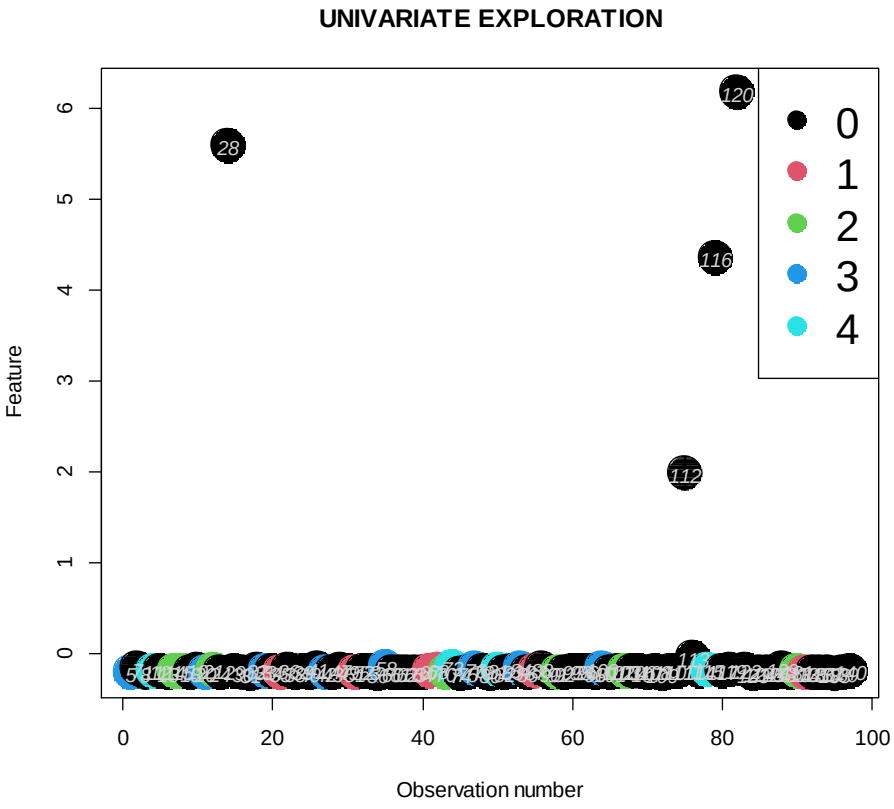


Ilustración 14 Distribución de observaciones en la variable R2W_std

Para corroborar las anomalías que estamos observando, procedemos a observar los valores respecto a los demás en el archivo Excel, para facilitar la comprensión. En la imagen 15 vemos que la observación NT 28 se caracteriza por unos valores muy distintos a los que se observan en el resto de las observaciones con una dispersión de valores muy elevada. Parece ser que los vóxeles de imagen poseían valores muy cercanos a 0 y alrededor de 1000.

NT	R2W_mean	R2W_std	R2W_median	R2W_p25	R2W_p75
15	66,5637148	10,0267574	66,8908565	60,3929838	73,1567274
18	43,4503141	8,57481067	43,2403792	37,3887756	48,9339786
19	73,8064052	11,7401828	74,2667905	66,1872566	81,7350232
20	45,0175913	7,24202165	45,1242416	40,5148913	49,6332342
21	82,7658153	12,9803216	82,2058703	73,7901034	91,6669993
24	39,9048058	7,95494737	39,9733748	34,3154461	45,3245111
28	497,13587	1063,50609	0,00029476	1,61E-05	683,239717
29	33,5753832	9,13888292	33,482948	27,1358928	39,7572708

Ilustración 15 Extracto de la BBDD en la que observamos que la observación 28 aparece con unos valores bastante atípicos.

De la misma manera se puede ver a continuación el resto de las observaciones mencionadas anteriormente con valores bastante atípicos. Hace falta más información para sacar alguna conclusión más precisa. Pero ¿pueden ser pacientes con alto grado de enfermedad? La respuesta es que no. Con esto estamos presuponiendo que los niveles altos de R2W

se corresponden con grados altos de enfermedad, no obstante, todos los pacientes tienen una clase 0 de Scheuer (hierro).

NT	R2W_mean	R2W_std	R2W_median	R2W_p25	R2W_p75
112	164,910264	407,103022	77,7541697	47,6137532	134,404245
113	84,5493044	38,1742206	81,8468331	65,1161285	97,9080051
114	63,3865576	9,88176203	63,8610334	56,7858525	70,1942163
115	85,861632	13,2833289	87,0546855	78,548354	94,9330043
116	335,025854	837,903574	155,881745	106,505846	258,651761
117	64,7470349	10,3895854	65,0223156	57,7415436	72,046604
119	23,0733631	12,13015	21,1265564	15,1594489	28,3920804
120	642,636181	1170,95971	379,027947	241,463726	618,241766

Ilustración 16 Extracto de la BBDD del resto de observaciones atípicas de las variables R2W.

En vista de los valores que han tomado estas observaciones se ha decidido eliminarlas a posteriori antes de incurrir en el entrenamiento de los diversos modelos estadístico, a excepción de sPLS y sPLS-DA, debido a que en ellos podemos evaluar las observaciones extremas y, detectaríamos quizá, incluso otras que podríamos estar pasando por alto en todo el *pool* de variables.

6.2.1 Exploración de las variables categóricas

Para observar el número de observaciones por cada clase dentro de cada sustrato patológico se realiza un análisis con la función *summary()* que rápidamente muestra información relevante. Con esta información se puede observar el desbalanceo de las distintas clases. El cambio que motivó a la creación de la segunda fase de análisis se puede observar en la Fibrosis ISHAK de la Tabla 2, cuya inclusión de covariables médicas hace muchos casos completos y considerando el desbalanceo no se puede entrenar un algoritmo con tan pocas clases en la fase más avanzada de la enfermedad. Esto es especialmente útil en *Random Forest* como ya se ha mencionado anteriormente, y en múltiples ocasiones generaba un error, impidiendo el análisis de dicha variable. Por ello se decidieron eliminar las variables ISHAK y Deugniers, a causa del elevado número de clases y la escasa distribución de pacientes en cada una de ellas.

Para el caso de la variable Deugniers, cabe recordar que dicha escala va de los valores 0-32, por lo que es altamente impracticable para una base de datos de 150 pacientes en el estudio, y todavía peor para los 49-97 casos completos, según el número de variables elegidas.

Tabla 2 El número de observaciones correspondiente a cada sustrato patológico en base a su escala determinada.

Sustrato patológico	Número de observaciones por clase						
	0	1	2	3	4	5	6
Fibrosis score	26	17	26	16	8		
Fibrosis score dicotomica	69	24	41	22	6		
Fibrosis ISHAK*	11 (23)	12 (17)	8 (21)	8 (12)	5 (9)	3 (7)	1 (5)
Fibrosis ISHAK dicotomica*	45 (85)	4 (12)					
Inflamación Lobular	24	41	22	6			
Inflamación Lobular dicotomica	65	28					
Inflamación Portal	36	58					
Inflamación Batts Ludwig	28	24	17	14	10		
Inflamación Batts Ludwig dicotomica	69	24					
Esteatosis	42	15	19	21			
Esteatosis dicotomica	57	40					
Scheuer	72	6	7	8	4		
Scheuer dicotomica	72	25					
Deugniers**							

* La variable ISHAK pertenece a la fase 1, por tanto se muestran el número de observaciones con todas las variables (RM + covariables médicas) y entre paréntesis el número de observaciones con las variables de RM.
** La variable Deugniers está fuera de está escala y se eliminó para la fase dos debido a su impracticabilidad, las clases serían las siguientes: 0 = 16 (45), 1 = 8 (11), 3 = 9 (12), 4 = 4 (7), 6 = 2 (6), 7 = 5 (5), 9 = 2 (3), 10 = 0 (1), 12 = 2 (3), 18 = 0 (1).
Nota = Las columnas vacías corresponden al final de la escala de valores para cada variable respuesta categórica.

6.2.2 Exploración de las variables continuas

En la tabla 3 se pude observar que las variables tienen valores mínimos cercanos a 0% a excepción del CD45, que tiene un mínimo de 2%. Esto podría ser debido a que todos los pacientes incluidos en el estudio presentan cierto grado de enfermedad, y la inflamación es el signo inicial más evidente mientras el hígado comienza a enfermar. También se observan máximos de 13%-15% para todas las enfermedades a excepción del hierro, lo que nos hace pensar que unos niveles de hierro altos impiden el desarrollo de las funciones hepáticas en mayor medida a otros niveles en condiciones similares. La inexistencia de valores mayores para todos los casos puede deberse a que la enfermedad ya se encuentra en estado avanzado con un riesgo de mortalidad elevado.

Tabla 3 Exploración de la distribución de las variables continuas.

Sustrato patológico	Distribución de las variables					
	Mínimo	1r Cuartil	Mediana	Media	3r Cuartil	Máximo
CPA	0,3328	6,2065	8,1418	8,0770	10,0532	14,0400
FPA	0,7360	2,6970	5,7480	6,2800	9,8580	13,3430
CD45	2,0300	4,2820	5,3100	6,3310	7,5210	15,9810
HIERRO	0,2093	1,2048	1,4720	1,8112	2,0819	6,3830

6.3 Relación entre las variables respuesta categóricas y continuas

Existen diversos modos de medir el grado o los niveles de cada enfermedad de los distintos sustratos hepatopatológicos. Uno de ellos es a partir de la caracterización categórica basada en niveles semicuantitativos, y otro sería a partir de la medición cuantitativa de un nivel de moléculas de interés presente en un tejido biológico (obtenido a partir de la biopsia). De esta manera, podríamos relacionar los siguientes sustratos:

- Fibrosis score/ISHAK con el área proporcional del colágeno (CPA).
- Inflamación Lobular/Portal/Batts Ludwig / Avanzada con la proteína CD45.
- Esteatosis con el área proporcional de grasa (FPA).
- Scheuer/Deugniers con los niveles de hierro (% hierro).

Inicialmente, se había realizado un ANOVA[12] para saber si los niveles cuantitativos pertenecen a las distintas poblaciones cualitativas, no obstante, tras varios descubrimientos en la literatura, se ha optado por realizar el método no paramétrico de Kruskal-Wallis, ya que no podemos asumir la distribución de normalidad en los datos.

Los descubrimientos sugieren que la progresión de la enfermedad en las distintas hepatopatologías puede no ser lineal, por lo que asumir una correlación entre los niveles de, por ejemplo, Fibrosis y el área de colágeno, podría ser un error.

6.3.1 Fibrosis score frente a CPA

La representación gráfica de los niveles de colágeno distribuidos en las distintas categorías muestra una tendencia a poseer altos niveles de colágeno para clases de Fibrosis alta.

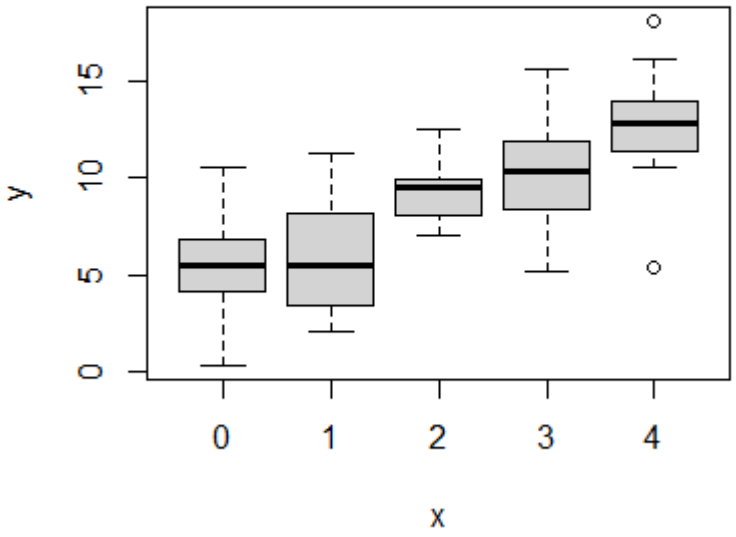


Ilustración 17 Cajas y bigotes de los niveles de CPA para las distintas clases de Fibrosis (score)

[12] No obstante, los resultados obtenidos demuestran un buen ajuste al papel probabilístico normal, homocedasticidad e independencia de los residuos, por lo que en principio podría utilizarse el ANOVA en la mayoría de los casos.

En el Anexo 4 en la prueba de Kruskal-Wallis podemos concluir que existen diferencias entre las distintas poblaciones, por lo que se rechaza la hipótesis nula de que las poblaciones son iguales.

Pese a poder encontrar diferencias estadísticamente significativas, y, apoyándonos en el gráfico de cajas y bigotes de la Imagen 17 se pueden visualizar cuales son las principales diferencias, se ha recurrido a la prueba de Wilcoxon para realizar una comparación entre pares (Anexo 5). Excepto en las categorías 0-1 y 2-3, existen diferencias significativas en el resto.

6.3.2 Fibrosis ISHAK frente a CPA

La Fibrosis ISHAK muestra también una relación positiva con los niveles de colágeno. Existe un elevado solapamiento, lo que dificulta la separación de clases tomando en cuenta solamente el área del colágeno.

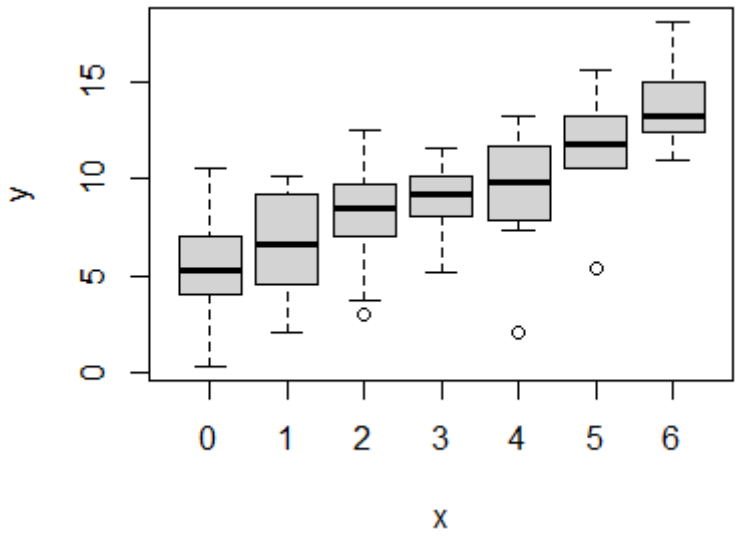


Ilustración 18 Cajas y bigotes de los niveles de CPA frente a las clases de Fibrosis ISHAK.

En el Anexo 6 podemos observar que se rechaza la hipótesis nula y aceptamos diferencias entre las distintas poblaciones. En la prueba de Wilcoxon podemos ver que no existen diferencias entre los grupos adyacentes 0-1, 1-2, 2-3, 3-4, 5-6, así como tampoco con el grupo 2-4 (Anexo 7).

6.3.3 Fibrosis score dicotomizado frente a CPA

Cuando dicotomizamos la Fibrosis score seguimos observando solapamiento, en la prueba de Kruskal-Wallis, en la que concluimos que existen diferencias significativas entre las dos poblaciones (Anexo 8). Como solo hay dos, no es necesario realizar comparación por pares mediante el método de Wilcoxon.

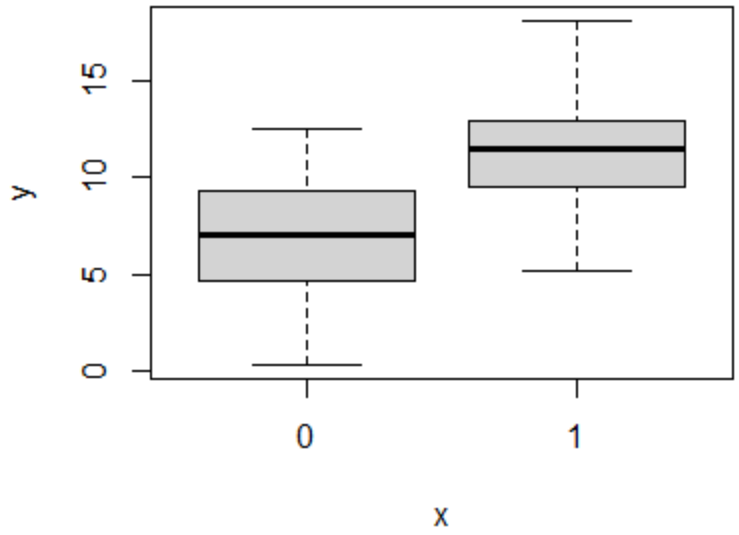


Ilustración 19 Cajas y bigotes de los niveles de CPA frente a las clases de Fibrosis Score Dicotomizado.

6.3.4 Fibrosis ISHAK dicotomizado frente a CPA

Se pueden apreciar diferencias significativas entre las dos clases (Anexo 9), con un grado de solapamiento algo menor, pero con una observación extremo con un valor bajo para el grado. Sería interesante observar que forma tomaría el gráfico con más observaciones, puesto que este sustrato patológico tenía un desbalanceo muy elevado para clases altas.

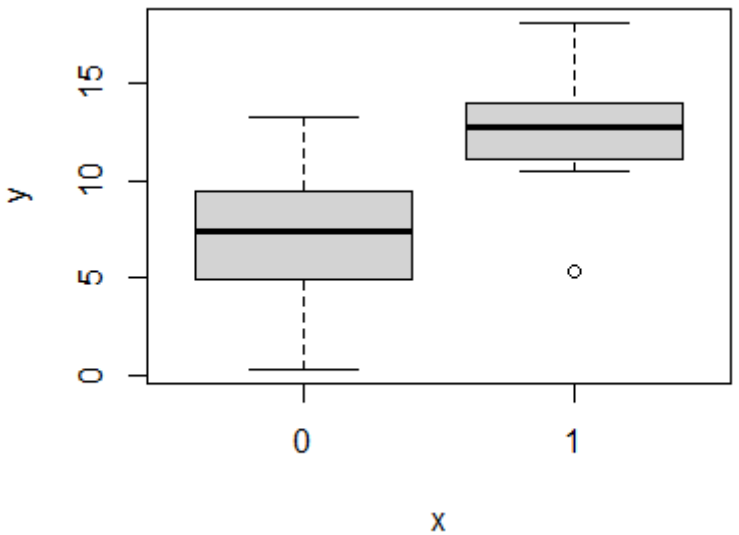


Ilustración 20 Cajas y Bigotes de Fibrosis ISHAK frente a CPA.

6.3.5 Inflamación Lobular frente a CD45

En este caso a nivel global aparecen diferencias significativas (Anexo 10), aunque en menor medida que en otros sustratos patológicos. En la comparación por pares solo aparecen diferencias significativas entre las clases 0-3 y 1-3 (Anexo 11).

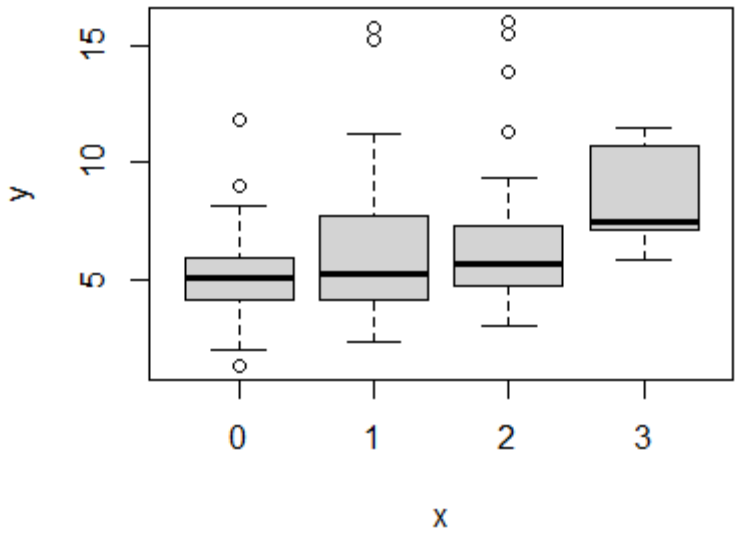


Ilustración 21 Cajas y bigotes de los niveles de CD45 frente a las clases de Inflamación Lobular.

6.3.6 Inflamación Portal frente a CD45

En la inflamación Portal podemos ver que la categoría ya está entre 0-1, por lo que nuevamente con la prueba de Kruskal-Wallis aparecen diferencias significativas entre las poblaciones (Anexo 12).

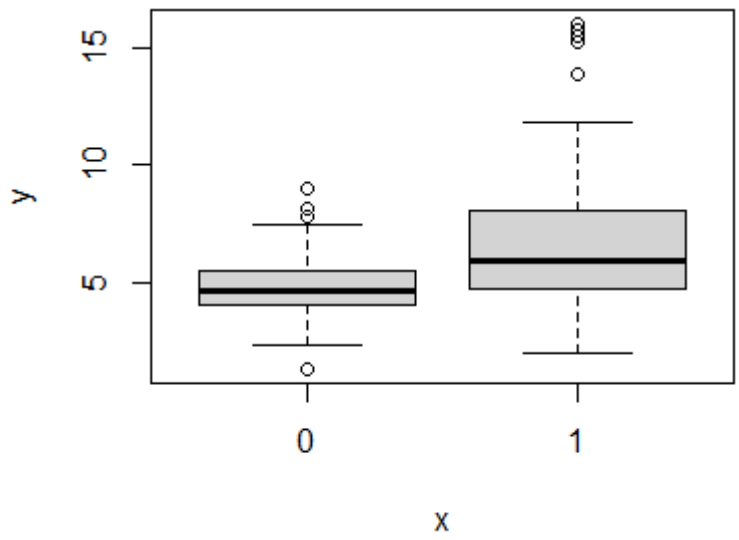


Ilustración 22 Cajas y bigotes de los niveles de CD45 frente a las clases de Inflamación Portal.

6.3.7 Inflamación Batts Ludwig frente a CD45

En la inflamación Batts Ludwig se aprecia una muy ligera variabilidad en los niveles de CD45 en las clases 1,2 y 3 con algunas observaciones más alejadas y en la clase 4 aparecen una elevada variabilidad que se solapa en gran medida con las otras clases. En la prueba de Kruskal-Wallis aparecen diferencias significativas (Anexo 13), y en la prueba de comparación por Wilcoxon las clases en las que se rechaza la hipótesis alternativa son las 0-1, 0-2, 1-2, 2-3, por lo pertenecerían a la misma población en términos estadísticos (Anexo 14).

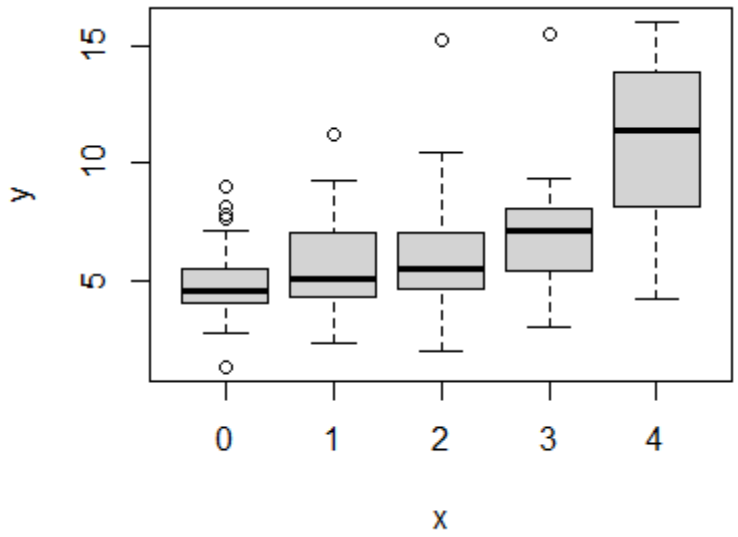


Ilustración 23 Cajas y bigotes de los niveles de CD45 frente a las clases de Inflamación Batts Ludwig.

6.3.8 Inflamación Lobular dicotomizada frente a CD45

De nuevo podemos ver un gran solapamiento, aunque existe una diferencia de medias significativa (Anexo 15).

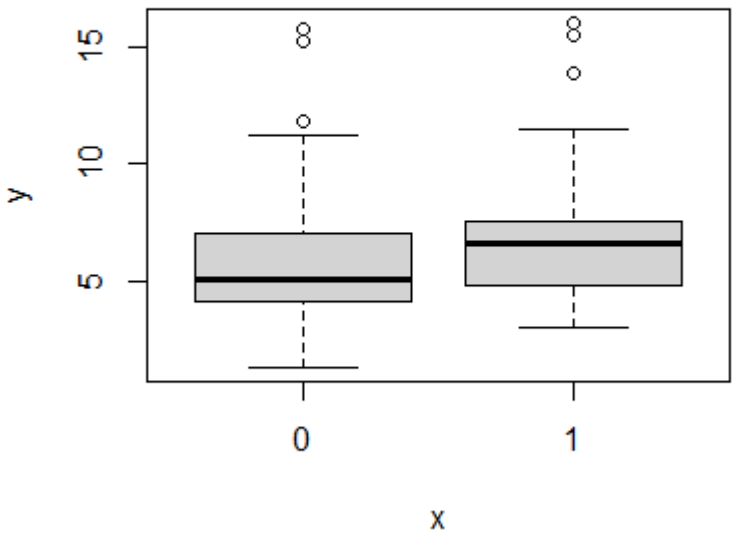


Ilustración 24 Cajas y Bigotes de los niveles de CD45 frente a las clases de inflamación Lobular dicotomizadas.

6.3.9 Inflamación Batts Ludwig dicotomizada frente a CD45

Tal y como se puede ver en el Anexo 16, la prueba de comparación de medias es estadísticamente significativo.

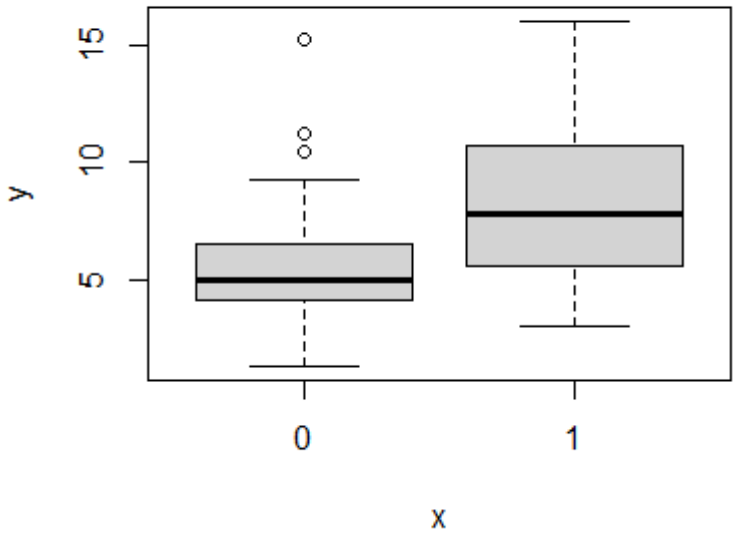


Ilustración 25 Cajas y Bigotes de los niveles de CD45 frente a las clases de inflamación Batts Ludwig dicotomizadas.

6.3.10 Esteatosis frente a FPA

En la Figura 20 podemos observar que debido a la clase 1 se forma un solapamiento completo, donde hay individuos con altos (y bajos) niveles de grasa clasificados con esta categoría, aunque a excepción de ello, se observa una tendencia alcista. En el Anexo 17 se concluyen diferencias significativas, aunque no sería así en las clases 1-2 (Anexo 18).

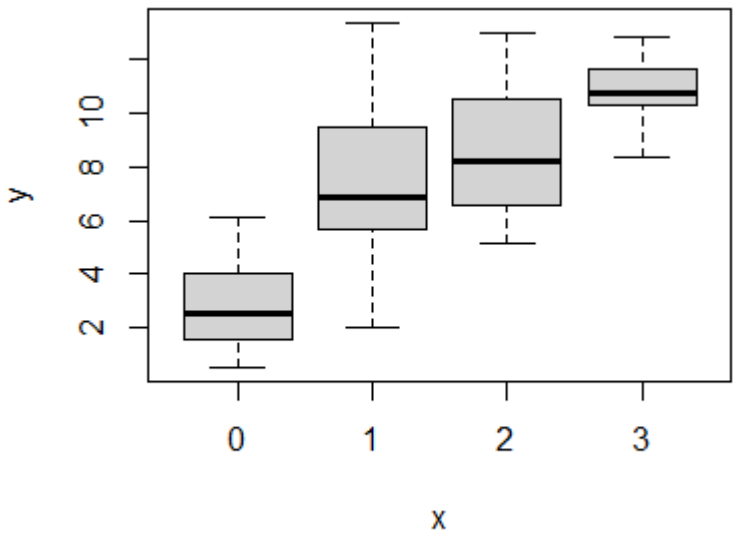


Ilustración 26 Cajas y Bigotes de los niveles de FPA frente a las clases de Esteatosis.

6.3.11 Esteatosis dicotomizada frente a FPA

Como podemos ver se aprecia una clara diferencia entre medias (Anexo 19). No obstante, sabiendo que estamos englobado el grado de Esteatosis 1 (apartado anterior) como clase 0 de esta dicotomización podemos ver que se mantiene gran parte de ese solapamiento.

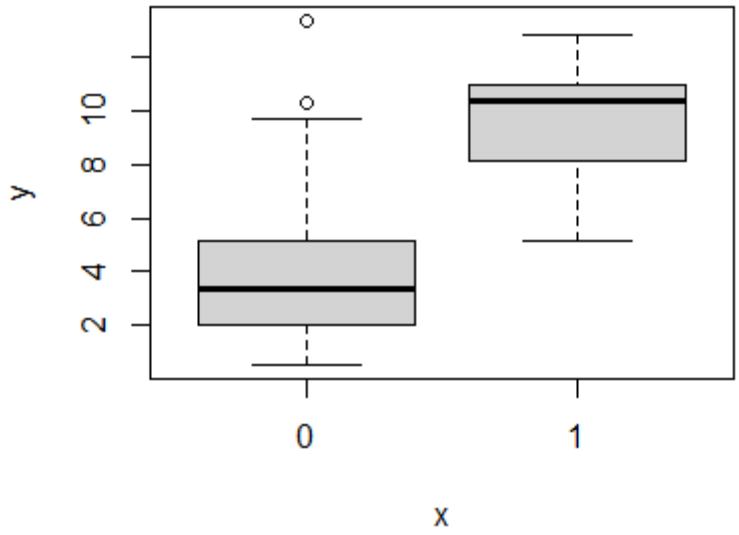


Ilustración 27 Cajas y Bigotes de los niveles de FPA frente a las clases de Esteatosis dicotomizada.

6.3.12 Deugniers frente al porcentaje de hierro total

Se observa que la clasificación asciende a medida que asciende el porcentaje de hierro total, con diferencias significativas entre clases (Anexo 20), concretamente entre las categorías 0-4, 0-6, 0-7, 0-9, 1-4, 1-7, 3-4, 3-7, 3-9 (Anexo 21).

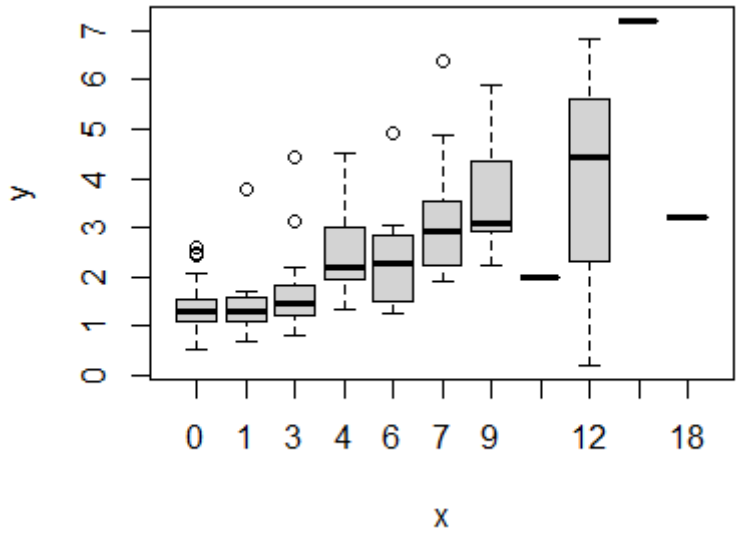


Ilustración 28 Cajas y Bigotes de los niveles de hierro frente a las clases de Deugniers.

6.3.13 Scheuer frente al porcentaje de hierro total

En este caso tan solo existen diferencias significativas (Anexo 22) del grupo 0 frente a las demás clases (Anexo 23).

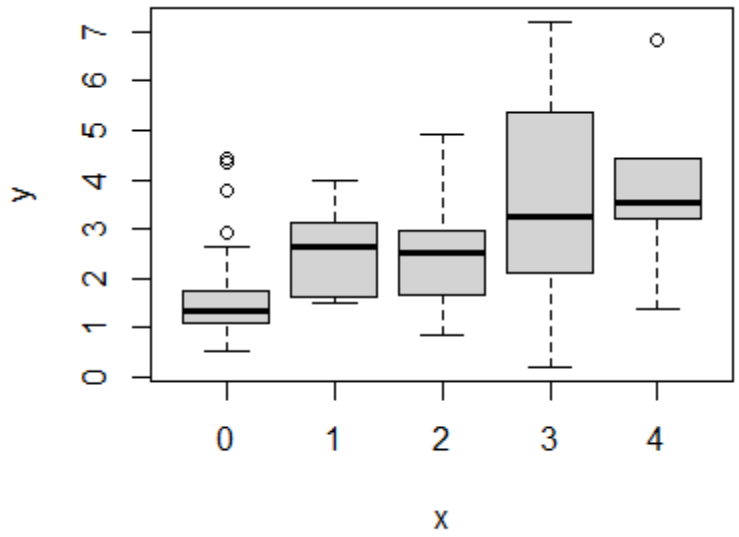


Ilustración 29 Cajas y Bigotes de los niveles de hierro frente a las clases de Scheuer.

6.3.14 Scheuer dicotómica frente al porcentaje de hierro total

Para este sustrato patológico existen diferencias significativas entre las dos poblaciones (Anexo 24).

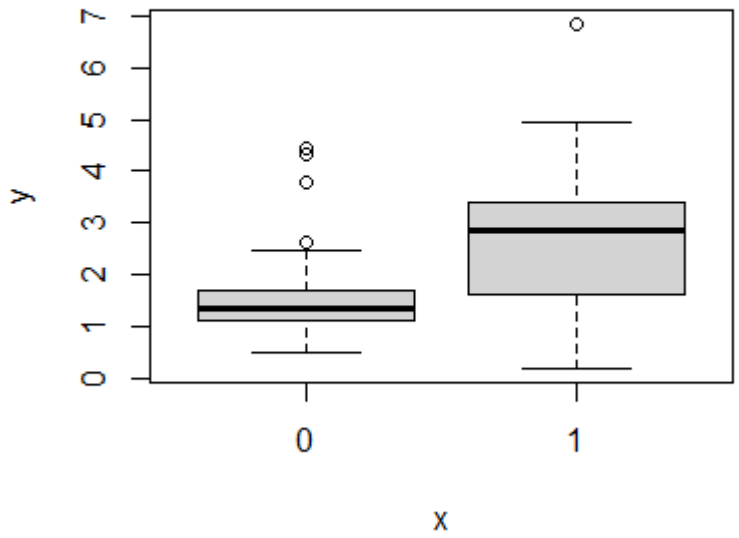


Ilustración 30 Cajas y Bigotes de los niveles de esteatosis dicotomizada frente al porcentaje de hierro total.

6.4 Análisis de las variables clínicas continuas

Los análisis de variables continuas han requerido de un estudio acerca de la distribución y el ajuste de los parámetros y hiperparámetros de los distintos modelos. En los modelos de SVM, kNN, *Random Forest* se han optimizado los hiperparámetros mediante *Grid Search*. Para la versión de regresión de LS-SVM los hiperparámetros se han optimizado

por Procesos Gaussinaos. Los parámetros de PCR, sPLS1 y sPLS2 (número de componentes principales) se han optimizado mediante *Leave one out* (Validación cruzada).

En cuanto a la selección de variables se ha utilizado el método de *backward stepwise* para la Regresión Lineal Múltiple y para PCR. Se ha utilizado regla 1-SE, que viene integrada en la librería de R para el algoritmo CART de Decision Tree. Para el modelo de *Random Forest* se han seleccionado las variables manualmente a través del gráfico de importancias (por reducción de la media y de Gini), así como del algoritmo VSURF. Para SVM y LS-SVM se ha utilizado el algoritmo RFE (basado en *Random Forest*) y para sPLS se ha utilizado una penalización Lasso (optimizada mediante validación cruzada *leave one out*).

Cabe mencionar que en los métodos estadísticos multivariantes como PLS y PCR permiten una optimización del número de componentes principales a utilizar a partir de validación cruzada (u otras técnicas de evaluación). No obstante, esto no es una selección de variables en sí misma, puesto que las variables latentes son una combinación lineal de las variables originales. Por tanto, elegir la componente principal con la que se realizará la predicción permite mejorar los resultados. Lo mismo se aplica para las variables categóricas a través de sPLS-DA y la regresión logística sobre las componentes principales. El paso en el que sí se seleccionan las variables es a través de la penalización ℓ1.

Se han realizado todos los análisis pertinentes al ajuste y validación de los modelos parámetricos del abanico de métodos empleados (análisis de residuos, R^2, $R^2_{ajustado}$, etc.), no obstante, al no haber ofrecido los mejores resultados no se incluirán en el TFM para no excederse en el volumen de contenidos.

Tabla 4 Resumen de resultados de las variables continuas.

SUSTRATO PATOLÓGICO	MÉTODO ESTADÍSTICO	(HIPER)PARÁMETROS OPTIMIZADOS	SELECCIÓN DE VARIABLES	PRECISIÓN DE OBUCHOWSKI
CPA	DECISION TREE	*	Nodo raíz: Ninguna	0,7172
CD45	KNN	K=3	**	0,7544
FPA	PCR***	ncomp = 2	Sin selección	0,7863
HIERRO	KNN	K=7	Sin selección	0,8275

* Post prunning (poda del árbol): regla de 1 – SE, del LIBRO CART (Breiman et. al. 1984)
**Variables KNN (CD45): f_p25, D^*_median, D^*_p75, D^*_p25, ADC_std_6b, IMC
***PCR equivale a *Principal Component Regression* o Regresión sobre componentes principales

Para las variables FPA y HIERRO se han repetido los análisis con la información que se obtuvo en el tiempo cronológico durante el cual se realizó la Fase 2. Eso afecta a la preselección de variables, ya que para FPA utilizamos las variables de resonancia magnética FF y para Hierro se utilizan las variables de resonancia R2W (con su media, mediana, desviación típica, y percentiles 25 y 75). La precisión de Obuchowski ha mejorado un 15,4567% y un 22,839%, respectivamente.

6.5 Análisis de las variables médicas categóricas

Siguiendo la línea de la metodología propuesta, tal y como hemos visto en las variables continuas, se ha realizado de la misma manera un profundo análisis iterativo, exploratorio y bibliográfico en materia de selección de variables, y por supuesto, del ajuste de hiperparámetros para las variables categóricas.

Los hiperparámetros se han ajustado de la siguiente manera. En *Random Forest*, SVM y kNN se han ajustado los hiperparámetros por Grid Search. En LS-SVM se han calculado mediante Procesos Gaussianos. Los parámetros del

número de componentes principales a elegir en Regresión Logística Binomial y Ordinal sobre componentes principales, y sPLS-DA se han calculado a partir del Error Cuadrático Medio (ECM o MSE en inglés).

En la selección de variables se ha utilizado la Penalización Lasso para la regresión logística ordinal a partir de la selección del valor lambda que produce un menor AIC (*Akaike Information Criteria* en inglés), y en su homólogo binomial de regresión logística, el valor de lambda se ha determinado por validación cruzada a partir del menor ECM. Para sPLS-DA se ha utilizado la penalización lasso a partir de un problema de optimización basado en la regla de Frobenius. Para SVM, kNN, LS-SVM la selección se ha basado en una primera de selección basado en *Random Forest* a partir del algoritmo RFE. Para *Random Forest* se ha realizado una selección manual basada en el gráfico de importancias y una selección automática mediante el algoritmo VSURF.

A continuación, se exponen los principales resultados de la investigación, cuya evaluación se centrará en la precisión de Obuchowski para las variables respuesta multiclase y en los parámetros de especificidad, sensibilidad y precisión para las variables respuesta binarias.

Tabla 5 Resumen de resultados de las variables categóricas.

SUSTRATO PATOLÓGICO	PARÁMETROS ÓPTIMOS								PRECISIÓN DE OBUCHOWSKI
	UMBRAL	TP	TN	FP	FN	ESPECIFICIDAD	SENSBILIDAD	PRECISIÓN	
FIBROSIS SCORE	-	-	-	-	-	-	-	-	0,9055
FIBROSIS AVANZADA	0,5901	50	19	18	4	0,7206	0,8695	0,7582	-
INFLAMACIÓN LOBULAR	-	-	-	-	-	-	-	-	0,8630
INF. LOBULAR AVANZADA	0,3863	52	11	13	17	0,7077	0,6786	0,6989	-
INFLAMACIÓN PORTAL	0,4626	21	50	15	8	0,5833	0,8793	0,766	-
INFLAMACIÓN BATTS LUDWIG	-	-	-	-	-	-	-	-	0,8441
INF. BATTS LUDWIG AVANZADA	0,3728	47	15	21	8	0,7206	0,6522	0,7033	-
ESTEATOSIS (GRASA)	-	-	-	-	-	-	-	-	0,9623
ESTEATOSIS AVANZADA (GRASA)	0,8423	54	35	2	6	0,9643	0,878	0,9278	-
SCHEUER (HIERRO)	-	-	-	-	-	-	-	-	0,8401
SCHEUER AVANZADO (HIERRO)	0,2773	23	27	17	17	0,5500	0,7273	0,6429	-

Adicionalmente, para aportar la información estadística que se complementa con los resultados de la Tabla 5, se ofrece una segunda tabla donde se expone el modelo estadístico utilizado, los parámetros o hiperparámetros seleccionados, las variables utilizadas en el modelo y las observaciones eliminadas para los casos donde se ha utilizado sPLS-DA (Tabla 6).

Las matrices de confusión se pueden consultar en los Anexos 25:35. Están coloreadas en función de los falsos negativos inadmisibles (rojo), falsos positivos inadmisibles (beige), diagonal de aciertos (verde) y dos regiones de azul claro que serían los falsos positivos y negativas que en la versión dicotomizada o avanzada de la enfermedad aparecerían agrupadas dentro el mismo valor 0 o 1.

Tabla 6 Información estadística subyacente a los resultados obtenidos en el análisis de variables categóricas.

SUSTRATO PATOLÓGICO	MÉTODO ESTADÍSTICO	(HIPER)PARÁMETROS	SELECCIÓN DE VARIABLES	OBSERVACIONES ELIMINADAS (NT)
FIBROSIS SCORE	*Random Forest*	mtry=4, ntree=4	Alpha_mean, D*_P75, Edad	-
FIBROSIS AVANZADA	*sPLS-DA*	CP=2, dist. centroide	*	113 y 122
INFLAMACIÓN LOBULAR	*Random Forest*	mtry=3, ntree=10	**	-
INF. LOBULAR AVANZADA	*Random Forest*	mtry=3, ntree=10	F_p25, Alpha_median, Alpha_std, ADC_median_2b	-
INFLAMACIÓN PORTAL	*Random Forest*	mtry=3, ntree=1	ratio_p75, Dalpha_p75, ratio_mean, T1_median	-
INFLAMACIÓN BATTS LUDWIG	*Random Forest*	mtry=2, ntree=10	***	-
INF. BATTS LUDWIG AVANZADA	*sPLS-DA*	CP=1, dist. centroide	Alpha_mean, Alpha_median	113 y 122
ESTEATOSIS (GRASA)	*Random Forest*	mtry=5, ntree=5	Todas (FF)	-
ESTEATOSIS AVANZADA (GRASA)	*Regresión Logística Ordinal (lasso)*	Lambda= 0.0129	FF_mean, FF_p25, FF_p75	-
SCHEUER (HIERRO)	*sPLS-DA*	CP=2, dist. centroide	R2W_std	28, 36, 58, 73, 75, 76, 84, 112, 113, 116, 119, 120, 131
SCHEUER AVANZADO (HIERRO)	*sPLS-DA*	CP=3, max.dist (euclídea)	R2W_p25, R2W_p75	36, 58, 73, 75, 76, 84, 110, 112, 113, 116, 120, 131

* Todas las variables a excepción de D_median, D_p25, kurt_mean, kurt_median, kurt_p25, kurt_p75, Alpha_p25
**T1_std, Ratio_p75, ADC_median_2b, Dalpha_p75, Alpha_std
***Dalpha_std, ADC_p75_6b, Alpha_median, Dalpha_p75, IMC, Dapp_std, Dapp_p75, Dapp_p25, Dapp_median, Alpha_mean, Ratio_mean

6.6 Desarrollo de las técnicas multivariantes

La técnica sPLS y sPLS-DA es especialmente útil para realizar un análisis basado en la comprensión de la estructura de datos que se está analizando. El principio de sparsity introducido nos puede ayudar además a seleccionar variables. Por otro lado, estas técnicas permiten observar la relación entre individuos y de variables.

La metodología que se ha seguido para el estudio de sustratos patológicos ha empezado por un análisis de los residuos del modelo a partir de un gráfico de residuos bidimensional de SPE y T2 de Hotelling, tal y como podemos ver en el Anexo 36, para la **Fibrosis Score dicotomizada**. Los residuos nos sirven para ver que observaciones han tomado valores desproporcionados en alguna variable. En este caso, las observaciones NT 122 y 113 poseen unos valores elevados en los residuos en SPE y T2, respectivamente. En el gráfico de *scores* (Ilustración 31) se puede ver como dichas observaciones nos quedan fuera de nuestro modelo. Esto provoca un efecto *leverage* que puede incidir en nuestros resultados. En el Anexo 37 se puede observar el mismo gráfico, pero con las observaciones eliminadas del modelo. Se pueden observar en ambos gráficos, con y sin eliminación de residuos, que la separación de clases no es buena. Mediante el gráfico de scores podemos observar las diferencias entre clases, y el cambio entre el Anexo 37 y la Ilustración 31 no parece demasiado sustancial, no obstante, estas observaciones pueden influir en las predicciones y en otras ocasiones, influir en otros gráficos, ofreciendo información errónea. Pese a que con el modelo PLS y PLS-DA tratamos de maximizar la covarianza de los datos, podemos observar como en la Ilustración 31 se explica el 27% y 12% y en el Anexo 37 estaríamos explicando el 28% y el 6% con las dos primeras componentes principales, respectivamente. Es fundamental modelizar correctamente con el procedimiento adecuado puesto que la información que se observa podría ser errónea.

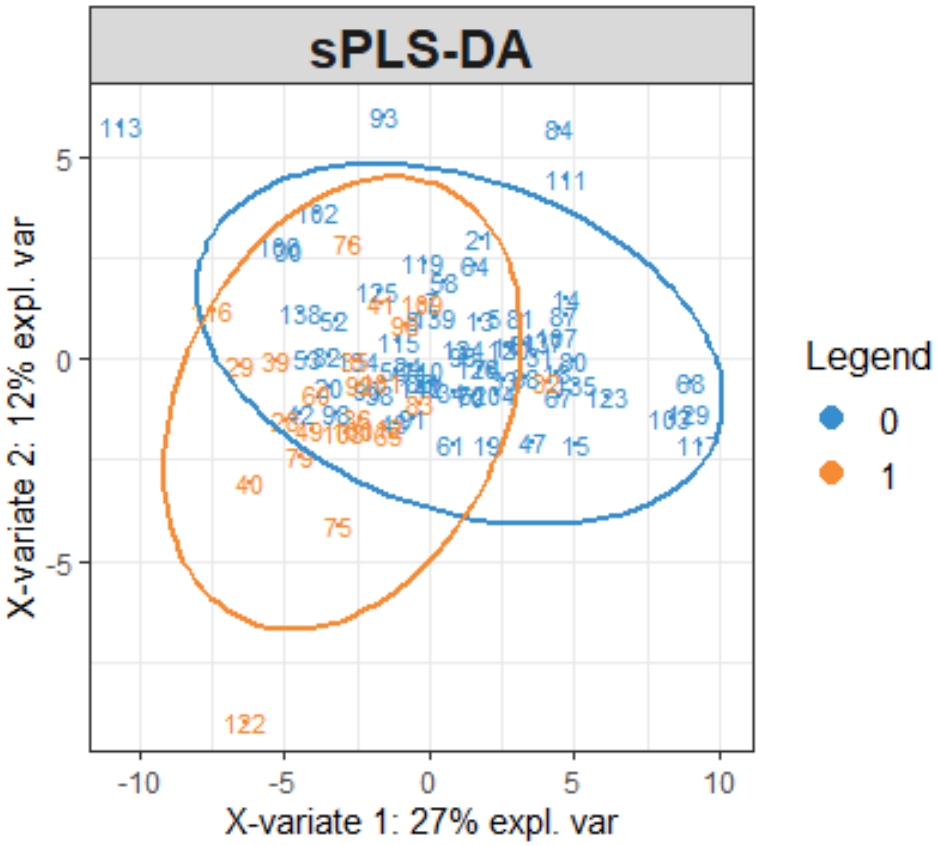


Ilustración 31 Gráfico de Scores sin eliminación de observaciones con residuos elevados. Fuente: Elaboración propia.

En cuanto las observaciones influyentes, la 122 tiene un valor elevado en la variable IMC (mayor peso en la segunda componente principal) y la 113 tiene un valor muy bajo de Dalpha_mean (peso negativo en la primera componente principal) y muy elevado en otras variables como cT1_mean, tal y como podemos ver en la Ilustración 32.

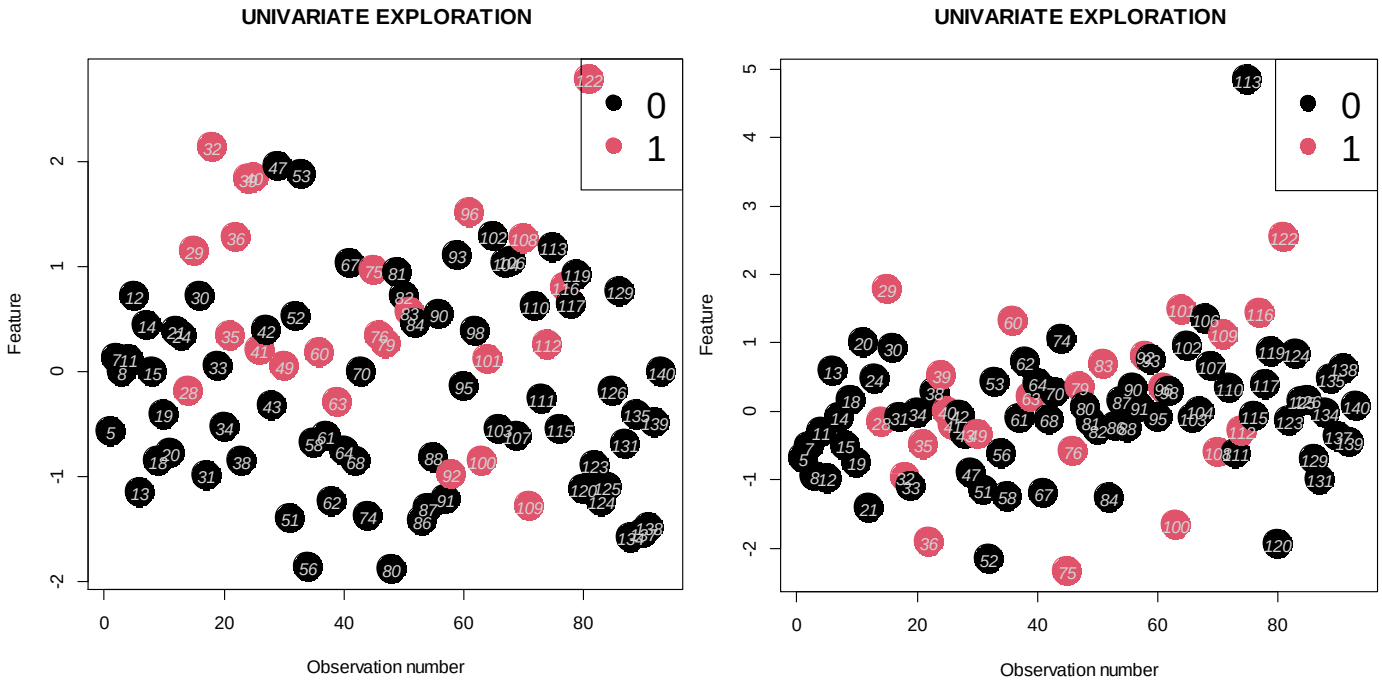


Ilustración 32 Variables IMC y cT1_mean, respectivamente, con las observaciones 122 y 113 algo por encima del resto de las observaciones, causando influencias extremas en el modelo. Fuente: elaboración propia.

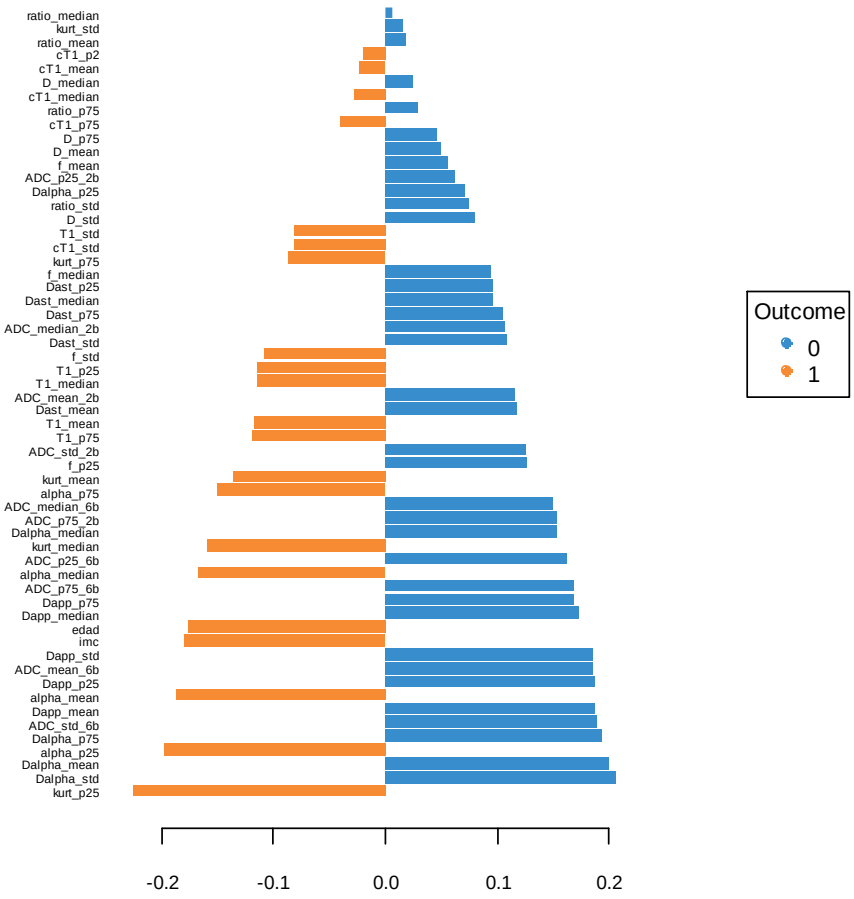


Ilustración 33 Gráfico de contribuciones de las variables en la primera componente principal. Fuente: elaboración propia.

En cuanto al modelo sin observaciones extremas, se recurre al gráfico de *loadings*, que nos muestra la relación entre las variables (Anexo 38). En él se puede observar un gran número de variables alrededor de 0 para los dos ejes, donde se asume que no tienen un gran efecto en la variable respuesta, y después existe una correlación negativa entre el grupo de variables Kurt y Alpha con el grupo ADC y Dalpha. Mediante el gráfico de contribuciones se puede observar el peso de cada variable por cada componente principal, además de la influencia para cada clase del sustrato patológico (Ilustración 33). Así, podemos observar que Kurt_p25 incide en el modelo con la predicción de la clase 1 y Dalpha_std con la predicción de la clase 0. De esta manera se visualiza una correlación negativa entre las distintas variables puesto que en la variable kurt_p25 los valores altos corresponden a una clase 0, y en cambio para Dalpha_std los valores altos corresponden con la clase 1 (Ilustración 34).

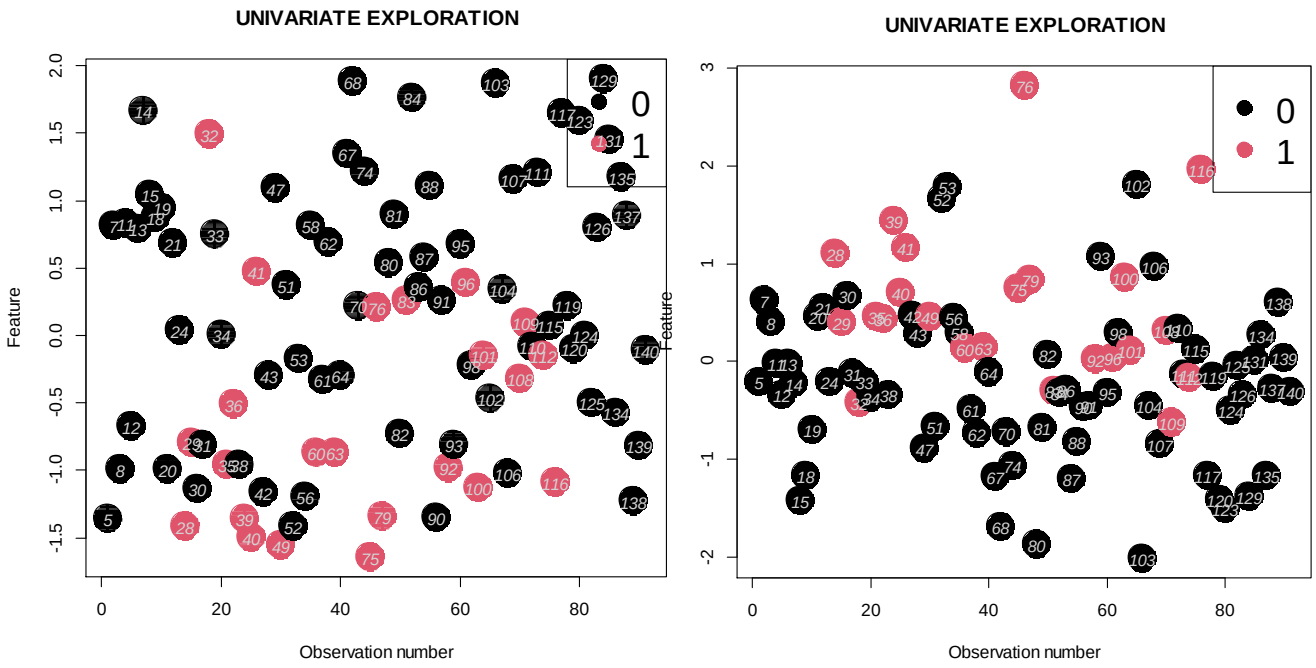


Ilustración 34 Gráficos Univariantes de las variables Kurt_p25 y Dalpha_std (centradas y escaladas), respectivamente. Fuente: Elaboración propia.

Observando los patrones correspondientes en los gráficos univariantes se propone la revisión de la clasificación de anatomía patológica para la observación NT 32 de Kurt_p25.

De manera análoga a la Fibrosis Score dicotomizada, para la **Inflamación Batts Ludwig** las observaciones NT 122 y 113 aparecían con unos residuos SPE yT2 elevados, por lo que se eliminaron del modelo (Anexo 39).

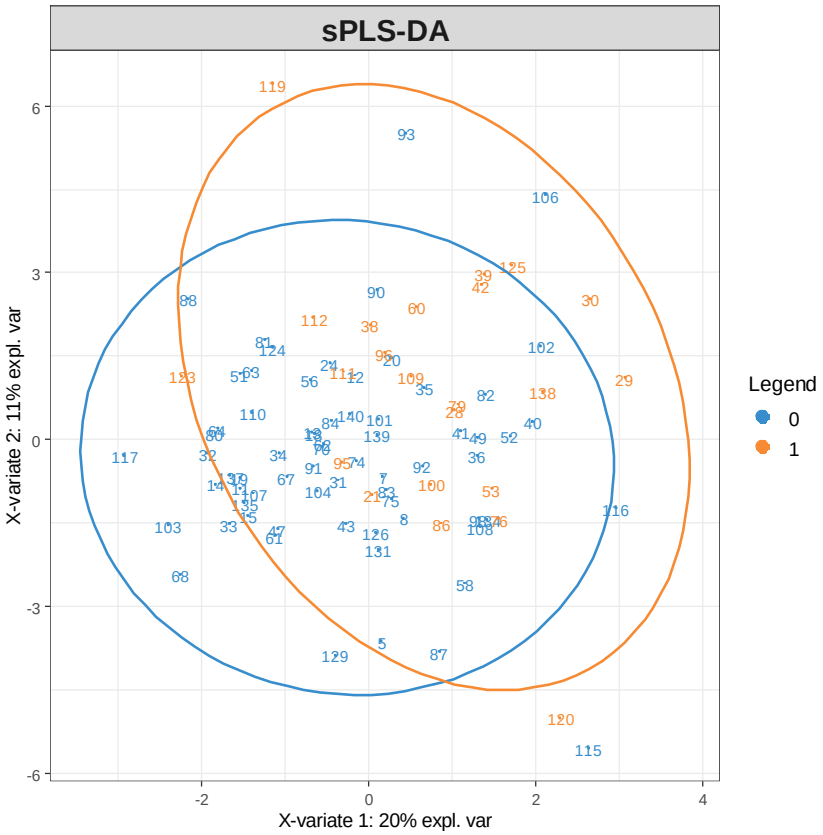


Ilustración 35 Gráfico de Scores del modelo sPLS-DA para la inflamación Batts Ludwig.

En el gráfico de *scores* (Ilustración 35) los pacientes aparecen muy solapados en cuanto a su clasificación, impidiendo una clara diferenciación de clases, y en los *loadings* se destacan las variables f_median y f_mean para la componente 1 y una lista más extensa para la componente 2 (Anexo 40). Desde el punto de vista anatomopatológico, se propone revisar las observaciones 84, 102, 105, 116 y 119, tal y como podemos ver en la variable f_median (Ilustración 36), que aparecen más alejadas y en la variable ratio_p25, las observaciones 81, 93 y 106 aparecen también alejadas.

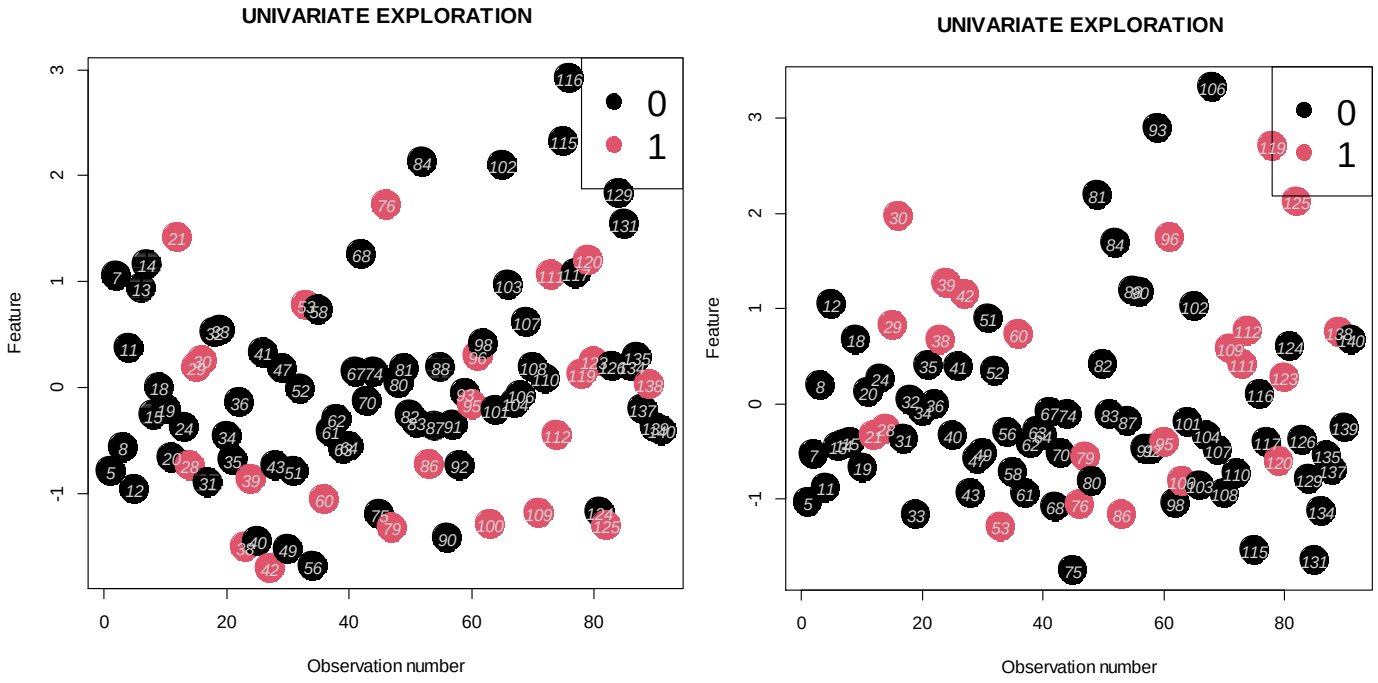


Ilustración 36 Variable f_median y ratio p_25para el grado de inflamación Batts Ludwig, respectivamente. Fuente: elaboración propia.

Finalmente, se propone realizar una revisión radiológica para entender por qué las observaciones 84, 5, 43 y 87 han tomado valores tan altos en la variable D_std (Ilustración 37).

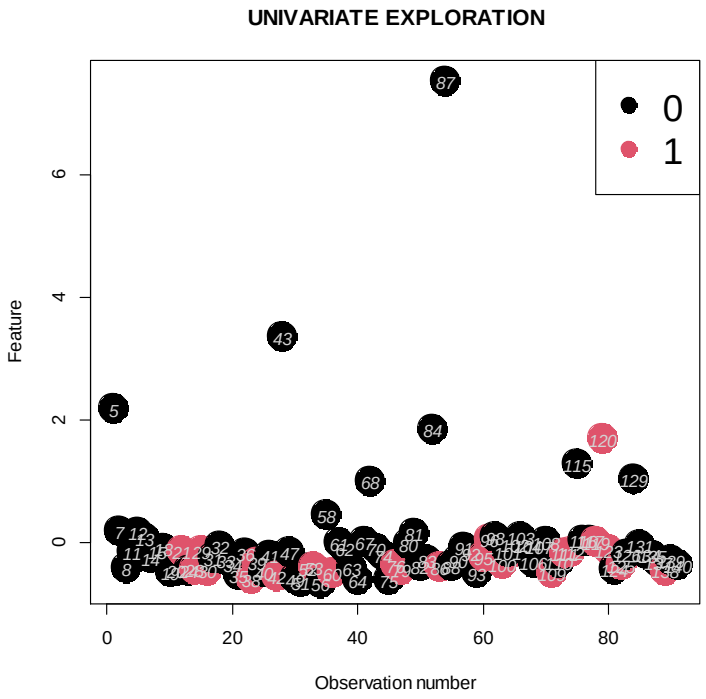


Ilustración 37 Variables ratio_p25 y D_std respectivamente con la etiqueta del grado de inflamación Batts Ludwig. Fuente: elaboración propia.

Para la **esteatosis**, poder analizar las variables más influyentes fue especialmente útil puesto que se detectaron dos observaciones que, tras una revisión y aprobación por un conjunto médico especialista, resultaron en un cambio de clase. Estas serían las observaciones 81 y 103 (Ilustración 38). Donde pasaron de una clasificación de 1 y 3 a 2 y 1, respectivamente.

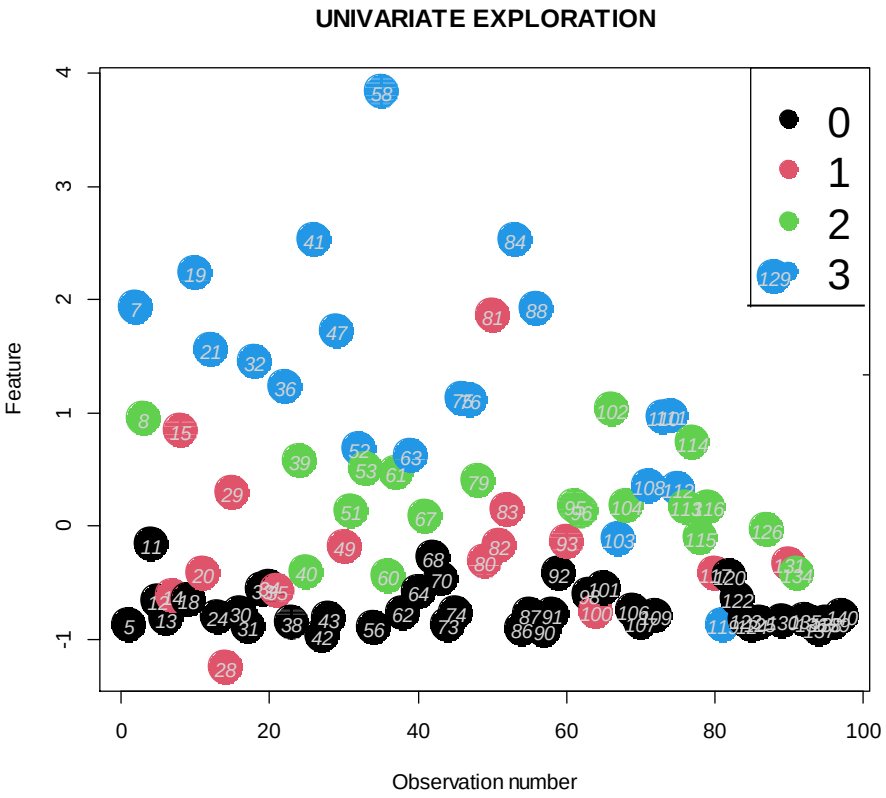


Ilustración 38 Variable FF_p75 para la etiqueta de Esteatosis 0-4. Fuente: elaboración propia.

En el gráfico de *scores* se puede observar cierto solapamiento, pero con mayor separabilidad respecto a los demás sustratos patológicos, con una varianza explicada del 80% para la primera componente y del 20% para la segunda (Ilustración 39).

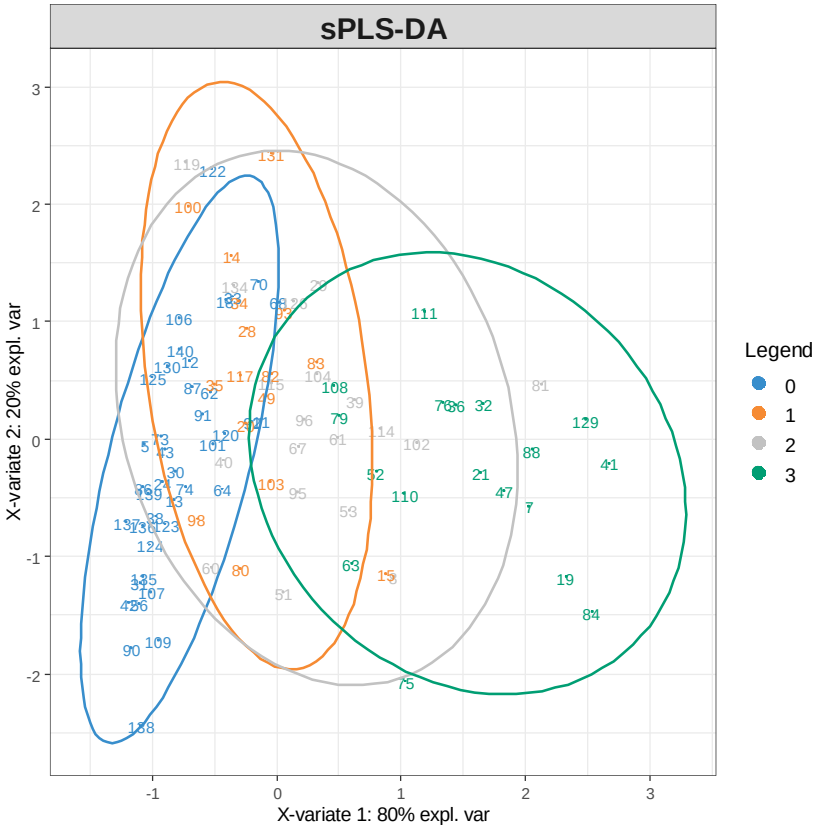


Ilustración 39 Gráfico de scores tras el cambio de clasificaciones en la observación 81 y 103 para Esteatosis. Fuente: elaboración propia.

Finalmente, para **Scheuer**, aparecen numerosas observaciones con unos residuos elevados, como se ha visto en la Tabla 5 de resultados, y se han tenido que eliminar del modelo (Anexo 41). No obstante, dichas observaciones requieren de un análisis radiológico para ver porqué tomaron valores tan altos. La relación entre individuos aparece con un alto grado de solapamiento y una varianza explicada del 100% para las dos primeras componentes (Anexo 42). Dichas observaciones son las mismas que se eliminan después en el modelo dicotomizado a excepción de la 28 (Anexo 43). La varianza explicada de las componentes 3 y 4 son de aproximadamente el 0%, no obstante, el modelo con 3 componentes nos genera el menor error en la predicción.

7 Conclusiones

El análisis entre los niveles cuantitativos de las variables relacionadas con su homólogo cualitativo permite ver la relación existente, en la que a mayor grado se observa un mayor nivel cuantitativo de las distintas variables. No obstante, se produce un alto grado de solapamiento entre las distintas clases debido a que las variables categóricas se basan, en criterios semicuantitativos, en ocasiones muy subjetivos y abiertos a la interpretación entre observadores, por lo que en ocasiones existen pacientes a los que se les ha diagnosticado como una clase 0 en algún sustrato hepatopatológico y sin embargo en la cuantificación de dicha variable se obtiene un valor que más bien parece corresponderse a un grado mayor. Esto es especialmente importante para la selección de nuevos criterios a la hora de realizar los análisis, que nos permitan medir la variable respuesta que se utilizará en futuras predicciones de manera más precisa, con menor variabilidad.

Se puede observar que las variables categóricas han ofrecido un mejor resultado en términos de precisión de Obuchowski respecto a las variables continuas pese a presentar el sesgo de introducir una mayor variabilidad en la clasificación de las observaciones por causas histológicas. No obstante, en las variables continuas se hace notable el

incremento de la precisión tras el cambio de variables de la fase 1 a la fase 2, especialmente en la grasa (esteatosis) y hierro hepático.

Los resultados en términos de Obuchowski oscilan en un intervalo de valores 0.72-0.96. El valor de predicción más alto lo ostenta la esteatosis (0'96), seguido de la fibrosis score (0'91), ambas variables categóricas. No existe un criterio de punto de corte a partir del cual se valore el cambio de referencia diagnóstica, y el objetivo más importante se centra en la fibrosis, que es la enfermedad más crítica, causando cirrosis en sus estadios más avanzados y elevando la mortalidad del paciente. Por tanto, se deben seguir realizando análisis para poder mejorar las predicciones. El método de Obuchowski funciona adecuadamente, además, para la validación de los modelos con categorías multiclase. La comparación de resultados con las variables dicotómicas está sujeta a sesgo, no obstante, la mejor predicción se realiza nuevamente con la esteatosis dicotómica, con un índice de Youden de 0'84, quedando la predicción de los demás sustratos patológicos por debajo de esta categoría. En cuanto a la precisión de Obuchowski de hierro es muy similar para la variable categórica scheuer y el porcentaje de hierro, aunque al dicotomizar el resultado ha empeorado significativamente, según su índice de Youden. En general, se puede apreciar una relación entre los biomarcadores (no invasivos) y los distintos sustratos patológicos estudiados, demostrando que las vías recorridas han sido acertadas en cuanto al uso de herramientas estadísticas para relacionar las imágenes de resonancia con las biopsias, permitiendo un amplio margen de mejora con el refinamiento de las técnicas de análisis de imagen y cuantitativas de biopsia.

En lo que respecta al descubrimiento de posibles biomarcadores, no se puede ofrecer una respuesta definitiva, y por tanto se debería seguir indagando en otras vías de investigación. No obstante, se sugiere la posibilidad de considerar las siguientes variables como posibles biomarcadores asociados a un sustrato patológico: D^* para CD45; Alpha y Kurt para Fibrosis Score y Score Avanzado; Alpha, Dapp, Dalpha para los distintos tipos de inflamación. En cuanto a los biomarcadores sugeridos por la literatura, a partir de las variables *Fat Fraction* (FF), se han obtenido muy buenos resultados en la predicción de la esteatosis, pero con R2W para predecir el hierro (Scheuer) no ha sido así.

En cuanto a las técnicas estadísticas que mejor resultado ha ofrecido el vecino más cercano (kNN) ha funcionado mejor en dos de los cuatro sustratos patológicos para las variables continuas, y *Random Forest*, seguido de sPLS-DA para las variables categóricas. En cuanto a la precisión de resultados se puede concluir que hay una clara relación entre las variables explicativas, no obstante, el hecho de que se hayan impuesto métodos no paramétricos podría sugerir relaciones no lineales con algunas de las variables que deberían investigarse, pudiendo incluso surgir cierta interacción entre ellas. Para el caso de las variables continuas, el equipo médico justificó que el algoritmo de imagen que calcula el área de elementos coloreados bajo la tinción histológica está en fase de pruebas y tienen un amplio margen de mejora. En cuanto a las variables categóricas, *Random Forest* crea reglas con las diferentes variables ofrecidas para tratar de hallar la solución que mejor se ajusta, sin los inconvenientes de sobreajuste que sí tiene *Decision Tree* y también emplea el algoritmo CART. Como incovenientes, con *Random Forest* no tenemos información de observaciones extremas, solo de variables menos importantes, por lo que su uso combinado con las herramientas estadísticas multivariantes ha sido bastante útil y, sobre todo, con un análisis exploratorio profundo.

8 Futuras Líneas

Se requieren investigaciones adicionales en la nosología de los criterios de las variables continuas que definen el estado del paciente como sano/enfermo. En este caso los resultados ofrecidos por el análisis de las variables continuas no se consideran aceptables, pero se debe seguir trabajando en los algoritmos de imagen que cuantifican los distintos niveles, para asociar de una manera más eficiente la relación entre las variables de resonancia magnética y los niveles de la enfermedad, a nivel cuantitativo. El objetivo sigue siendo, por tanto, seguir indagando y mejorando las distintas herramientas que nos lleven a dejar de biopsiar a los pacientes con un alto grado de precisión en el diagnóstico.

Por otro lado, las imágenes de la resonancia magnética deben poder abarcar el hígado en su totalidad para caracterizar las distintas enfermedades que se han visto a lo largo del TFM, ya que actualmente se produce un sesgo debido a que se estudian regiones de interés donde se ha observado que los depósitos de colágeno, grasa, hierro, etc., se acumulan con mayor frecuencia en zonas como la NASH. Así mismo, el futuro de la medicina clusterizará el hígado en distintas secciones para mostrar cuales han sido las regiones más afectadas, y así poder determinar el punto en el que se debe de operar.

9 Bibliografía

A.Cequera, & Méndez, M. d. (2014). Biomarcadores para fibrosis hepática, avances, ventajas y desventajas. *Revista de Gastroenterología de México*, Volume 79, Issue 3, July–September, Pages 187-199.

Afshin, M., Sadeghian, A., & Raahemifar, K. (2007). On Efficient Tuning of LS-SVM Hyper-Parameters in Short-Term Load Forecasting: A Comparative Study. *IEEE Power Engineering Society General Meeting*, (págs. 1-6). doi:10.1109/PES.2007.385613

App4Stats. (s.f.). Obtenido de http://app4stats.com/capitulo-vii-2/

Archer, K. J., & Williams, A. A. (2012). L1 Penalized Continuation Ratio Models for Ordinal Response Prediction Using High-dimensional Datasets. *Statistics in Medicine, 31*, 1464-1474. Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3718008/

Asrani, S., Devarbhavi, H., Eaton, J., & Kamath, P. S. (January de 2019). Burden of liver diseases in the world. *J Hepatol, 70*(1), 151-171.

Audette, M. A., Ferrie, F. P., & Peters, T. M. (2000). An algorithmic overview of surface registration techniques for medical imaging. *Medical image analysis*, 201-217.

Bender, R., & Grouven, U. (1997). Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London, 31*(5), 546.

Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics, 7*(1), 142.

Breiman, L. (2004). Consistency for a simple model of random forests.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees. Statistics/probability series.* California, USA: Wadsworth Publishing Company.

Brochu, E., Vlad, M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning.

Chevallier, D. M., Guerret, S., Chossegros, P., Gerard, F., & Grimaud, J.-A. (1994). A histological semiquantitative scoring system for evaluation of hepatic fibrosis in needle liver biopsy specimens: Comparison with morphometric studies. *AASLD*, Volume 20, Issue 2.

Chung, D., Chun, H., & Keles, S. (2012). Spls: sparse partial least squares (SPLS) regression and classification. *R package, version, 2*, 1-1.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cortez, P. (2020). rminer: Data Mining Classification and Regression Methods. Obtenido de https://CRAN.R-project.org/package=rminer

Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics, 19*(1), 65.

Dawant, B. M. (2002). Non-rigid registration of medical images: purpose and methods, a short survey. *Proceedings IEEE International Symposium on Biomedical Imaging*, 465-468.

Dolan, D. H., & Jones, S. C. (2008). THRIVE: a data reduction program for three-phase PDV/PDI and VISAR measurements. *Sandia National Laboratories*, SAND2008-3871.

Donato, H., França, M., Candelária, I., & Caseiro-Alves, F. (2017). Liver MRI: From basic protocol to advanced techniques. *European Journal of Radiology, 93*, 30-39.

Douzas, G. B. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences, 465*, 1-20.

Drozdowicz, B., Bernasconi, G., Reyes, M., Saba, F., & Simón, G. (2005). Segmentación semiautomática de imágenes de resonancia magnética, basada en redes neuronales artificiales. *Ciencia, Docencia y Tecnología*, 117-155.

Dunn, K. (2020). *Process Improvement Using Data.* Obtenido de https://learnche.org/pid/PID.pdf?76ba30

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

Fattovich, G., Giustina, G., Degos, F., Tremolada, F., Diodati, G., Almasio, G., . . . Brouwer., J. (February de 1997). Morbidity and mortality in compensated cirrhosis type C: a retrospective follow-up study of 384 patients. *Gastroenterology, 112*(2), 463-72.

Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. En *Advances in neural information processing systems* (págs. 838-846).

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 47*(4), 458-472.

Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression. (Third Edition). Thousand Oaks CA. Obtenido de https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via. *Journal of Statistical Software, 33*(1), 1-22.

Friedman, S. L. (JANUARY de 2003). Liver fibrosis – from bench to bedside. *Journal of Hepatology, 38*(01), 38-53.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta, 185*, 1-17.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: an R package for variable selection using random forests. *The R Journal, 7*(2), 19-33. doi:hal-01251924

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2019). VSURF: Variable Selection Using Random Forests. Obtenido de https://CRAN.R-project.org/package=VSURF

Gey, S., & Nedelec, E. (2005). Model selection for CART regression trees. *IEEE Transactions on Information Theory, 51*(2), 658-670.

Girden, E. R. (1992). ANOVA: Repeated measures. (84).

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics, 31*(15), 2595-2597.

Grigorescu, M. (2006). Noninvasive biochemical markers of liver fibrosis. *Journal of Gastrointestinal and Liver Diseases : JGLD*, 15(2):149-159.

Hagström, H., Nasr, P., Ekstedt, M., Stal, P., Hultcrantz, R., & Kechagias, S. (2017). Fibrosis stage but not NASH predicts mortality andtime to develpment of severe liver diseas in biopsy-proven NAFLD. *Journal of Hepatology, 67*, 1265-1273. doi:http://dx.doi.org/10.1016/j.jhep.2017.07.027

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Hatta, T., Fujinaga, Y., Kadoya, M., Ueda, H., Murayama, H., Kurozumi, M., . . . Tanaka, N. (2010). Accurate and simple method for quantification of hepatic fat content using magnetic resonance imaging: a prospective study in biopsy-proven nonalcoholic fatty liver disease. *Journal of gastroenterology*, 1263-1271.

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods, 9*(10), 1043-1069. doi:10.1080/03610928008827941

Karatzoglou, A., Smola, A., Hornik, K., & Zieleis, A. (2004). An S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*(9), 1-20. Obtenido de http://www.jstatsoft.org/v11/i09/

Kemalbay, G., & Korkmazoğlu, Ö. B. (2014). Categorical principal component logistic regression: a case study for housing loan approval. *Procedia-Social and Behavioral Sciences, 109*, 730-736.

Kershenobich, D. S., & Weissbrod, A. B. (2003). Liver fibrosis and inflammation. A review. *Annals of hepatology, 2*(4), 159–163.

Kervrann, C., Boulanger, J., & Coupé, P. (2007). Bayesian non-local means filter, image redundancy and adaptive dictionaries for noise removal. *International conference on scale space and variational methods in computer vision*, 520-532.

Kleiner, D. E., Brunt, E. M., Natta, M. V., Behling, C., Contos, M. J., Cummings, O. W., . . . Sanyal, A. J. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*, 1313-1321.

Kucheryavskiy, S. (2020). mdatools - R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems, 198*. Obtenido de https://doi.org/10.1016/j.chemolab.2020.103937

Kuhn, M. (2020). caret: Classification and Regression Training. Obtenido de https://CRAN.R-project.org/package=caret

Kulesa, A., Krzywinski, M., Blainey, P., & Altman, N. (28 de Mayo de 2015). Sampling distributions and the bootstrap. *Nature Methods, 12*, 477-48. doi:https://doi.org/10.1038/nmeth.3414

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevante feature selection and graphical displays for multiclass problems. *BMC Bioinformatics, 12*(1), 253. Obtenido de http://www.biomedcentral.com/1471-2105/12/253

Lee, L. C., Liong, C.-Y., & Jemain, A. A. (2018). Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst, 143*(15), 3526-3529.

Lerman, P. (1980). Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society: Series C (Applied Statistics, 29*(1), 77-84.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22. Obtenido de https://CRAN.R-project.org/doc/Rnews/

Manuela França, Á. A.-B.-B. (2017). Accurate simultaneous quantification of liver steatosis and iron overload in diffuse liver diseases with MRI. *Abdominal Radiology, 42*(5), 1434-1443. doi:10.1007/s00261-017-1048-0

Martí Aguado, D., Rodríguez Ortega, A., Alagarda Mestre, C., Bauza, M., Valero Pérez, E., Alfaro Cervello, C., . . . Martí Bonmati, L. (2020). Digital pathology: accurate technique for quantitative assessment of histological features in metabolic-associated fatty liver disease. *Alimentary Pharmacology & Therapeutics*.

Marti-Aguado, D., Rodriguez-Ortega, A., Carot, J. M., Mestre-Alagarda, C., Bauza, M., Valero-Perez, E., . . . Martí-Bonmatí, L. (s.f.). *Multiparametric MR analysis of liver fibrosis and inflammation: from gray zones to clinical relevance*. *Abdominal Radiology*. En revisión.

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1-1.

Medina, F., & Galván, M. (2007). *Imputación de datos: teoría y práctica.* Cepal.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability. Obtenido de https://CRAN.R-project.org/package=e1071

Microsoft Corporation. (2019). Microsoft Excel. Obtenido de https://office.microsoft.com/excel

Molloy, J. W., Calcagno, C. J., Williams, C. D., Jones, F. J., Torres, D. M., & Harrison, S. A. (2011). Association of coffee and caffeine consumption with fatty liver disease, nonalcoholic steatohepatitis, and degree of hepatic fibrosis. *AASLD*, Volume 55, Issue 2.

Nasr, P., Fredrikson, M., Ekstedt, M., & Kechagias, S. (2020). The amount of liver fat predicts mortality and development of type 2 diabetes in non-alcoholic fatty liver disease. *Liver International, 40*(5), 1069-1078.

Nelson, J., Wilson, L., Brunt, E. M., Yeh, M. M., Kleiner, D. E., Unalp-Arida, A., . . . NASH CRN. (February de 2011). Relationship between pattern of hepatic iron deposition and histologic severity in nonalcoholic fatty liver disease. *Hepatology, 53*(2), 448-457. doi:doi:10.1002/hep.24038

Nguyen, P. (2012). *nonbinROC: ROC-type analysis for non-binary gold standards.* Obtenido de https://CRAN.R-project.org/package=nonbinROC

Obuchowski, N. A. (2005). Estimating and Comparing Diagnostic Tests. *Academic radiology, 12*(9), 1198-1204.

Obuchowski, N. A. (2005). Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Academic radiology, 12*(9), 1198-1204.

Obuchowski, N. A., Goske, M. J., & Applegate, K. E. (2001). Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. *Statistics in medicine, 20*(21), 3261-3278.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia, 4*(2), 1883.

R Core Team. (2020). R. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Obtenido de https://www.R-project.org/

Ratziu, V., Charlotte, F., Heurtier, A., Gombert, S., Giral, P., Bruckert, E., . . . Poynard, T. (2005). Sampling Variability of Liver Biopsy in Nonalcoholic Fatty Liver Disease. *Gastroenterology*, 1898-1906.

Rifai, K., Cornberg, J., Mederacke, I., Bahr, M. J., Wedemeyer, H., Malinski, P., & Gebel, M. (2011). Clinical feasibility of liver elastography by acoustic radiation force impulse imaging (ARFI). *Digestive and Liver Disease*, 43(6), 491-497.

Rodrigo, J. A. (Abril de 2018). Machine Learning con R y caret. Obtenido de https://rpubs.com/Joaquin_AR/383283

Rohart, F., Gautier, B., Singh, A., & Cao, K.-A. L. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology, 13*(11), e1005752. Obtenido de http://www.mixOmics.org

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K. A. (2017). mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS computational biology, 13*(11), e1005752.

RStudio Team. (2020). RStudio: Integrated Development Environment for R. Boston, MA. Obtenido de http://www.rstudio.com/

Snedecor, G. W., & Cochran, W. (1967). Statistical methods 6th edition. *The Iowa State University*.

Thiele, C., & Hirschfeld, G. (2020). cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R. *arXiv*.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology, 73*(3), 273-282.

Torgo, L. (2010). *Data Mining with R, learning with case studies.* Chapman and Hall/CRC. Obtenido de http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

Vieira, D. C., Neto, A. R., & Rodrigues, A. W. (2016). Sparse least squares support vector regression via multiresponse sparse regression. *2016 International Joint Conference on Neural Networks (IJCNN)* (págs. 3218-3225). IEEE.

Wandishin, M. S., & Mullen, S. J. (2009). Multiclass ROC analysis. *Weather and Forecasting, 24*(2), 530-547.

Wickham, H., & Bryan, J. (2019). readxl: Read Excel Files. (R package version 1.3.1). Obtenido de https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. (R package version 1.0.2). Obtenido de https://CRAN.R-project.org/package=dplyr

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. En *International conference on machine learning* (págs. 1067-1075).

Woolson, R. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, 1-3.

Worley, B., Halouska, S., & Powers, R. (2013). Utilities for quantifying separation in PCA/PLS-DA scores plots. *Analytical biochemistry, 433*(2), 102-104.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology, 17*(1), 26-40.

10 Anexos

A continuación, se adjuntan las tablas e imágenes más relevantes para apoyar las conclusiones de este TFM.

Anexo 1 Matriz de penalización para los sustratos patológicos con 4 clases (0-3).

Matriz de penalización	0	1	2	3
0	0	0.25	0.50	0.75
1	0	0	0.25	0.5
2	0	0	0	0.25
3	0	0	0	0

Anexo 2 Matriz de penalización para los sustratos patológicos con 5 clases (0-4).

Matriz de penalización	0	1	2	3	4
0	0	0.25	0.50	0.75	1
1	0	0	0.25	0.50	0.75
2	0	0	0	0.25	0.50
3	0	0	0	0	0.25
4	0	0	0	0	0

Anexo 3 Matriz de penalización para los sustratos patológicos con 7 clases (0-6). Concretamente utilizada para la Fibrosis ISHAK.

Matriz de penalización	0	1	2	3	4	5	6
0	0	0.15	0.30	0.45	0.60	0.75	1
1	0	0	0.15	0.30	0.45	0.60	0.75
2	0	0	0	0.15	0.30	0.45	0.60
3	0	0	0	0	0.15	0.30	0.45
4	0	0	0	0	0	0.15	0.30
5	0	0	0	0	0	0	0.15
6	0	0	0	0	0	0	0

Anexo 4 Test de Kruskal-Wallis para Fibrosis Score y CPA.

```
        Kruskal-Wallis rank sum test

data:  cpa.model.1$cpa by cpa.model.1$fibscore
Kruskal-Wallis chi-squared = 73.473, df = 4, p-value = 4.191e-15
```

Anexo 5 Comparación entre categorías de Fibrosis Score mediante la prueba de Wilcoxon.

```
        Pairwise comparisons using Wilcoxon rank sum exact test

data:  cpa.model.1$cpa and cpa.model.1$fibscore

  0       1       2       3
1 0.960   -       -       -
2 7.8e-09 2.4e-05 -       -
3 5.4e-09 2.3e-06 0.069   -
4 8.5e-09 1.3e-07 2.4e-05 0.017
```

Anexo 6 Test de Kruskal-Wallis para Fibrosis ISHAK y CPA

```
        Kruskal-Wallis rank sum test

data:  cpa.model.2$cpa by cpa.model.2$fibishak
Kruskal-Wallis chi-squared = 63.459, df = 6, p-value = 8.897e-12
```

Anexo 7 Comparación entre categorías mediante la prueba de Wilcoxon entre Fibrosis ISHAK y CPA.

```
        Pairwise comparisons using Wilcoxon rank sum exact test

data:  cpa.model.2$cpa and cpa.model.2$fibishak

  0       1       2       3       4       5
1 0.19558 -       -       -       -       -
2 0.00064 0.07627 -       -       -       -
3 1.9e-05 0.00659 0.26249 -       -       -
4 0.00031 0.00599 0.15839 0.45586 -       -
5 4.2e-07 1.8e-06 0.00024 0.00037 0.03081 -
6 6.5e-07 2.0e-06 0.00032 8.2e-05 0.00603 0.15839
```

Anexo 8 Test de Kruskal-Wallis para Fibrosis Score dicotomizado y CPA.

```
        Kruskal-Wallis rank sum test

data:  cpa.model.3$cpa by cpa.model.3$fibdic
Kruskal-Wallis chi-squared = 43.71, df = 1, p-value = 3.808e-11
```

Anexo 9 Test de Kruskal-Wallis para Fibrosis ISHAK dicotomizado y CPA

```
        Kruskal-Wallis rank sum test

data:  cpa.model.4$cpa by cpa.model.4$fibishakdic
Kruskal-Wallis chi-squared = 38.896, df = 1, p-value = 4.469e-10
```

Anexo 10 Test de Kruskal-Wallis para Inflamación Lobular y CD45.

```
        Kruskal-Wallis rank sum test

data:  cd45.model.1$cd45 by cd45.model.1$lobular
Kruskal-Wallis chi-squared = 11.054, df = 3, p-value = 0.01144
```

Anexo 11 Test de Kruskal-Wallis para Inflamación Portal y CD45.

```
        Kruskal-Wallis rank sum test

data:  cd45.model.2$cd45 by cd45.model.2$portal
Kruskal-Wallis chi-squared = 15.689, df = 1, p-value = 7.467e-05
```

Anexo 12 Test de Kruskal-Wallis para Inflamación Batts Ludwig y CD45.

```
        Kruskal-Wallis rank sum test

data:  cd45.model.3$cd45 by cd45.model.3$battsludwig
Kruskal-Wallis chi-squared = 30.207, df = 4, p-value = 4.442e-06
```

Anexo 13 Comparación entre categorías mediante la prueba de Wilcoxon entre Inflamación Batts Ludwig y CD45.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  cd45.model.3$cd45 and cd45.model.3$battsludwig

  0       1       2       3
1 0.11438 -       -       -
2 0.05711 0.60496 -       -
3 0.00069 0.02289 0.10862 -
4 2.5e-05 0.00016 0.00131 0.00688
```

Anexo 14 Test de Kruskal-Wallis para Inflamación Lobular dicotomizada y CD45.

```
        Kruskal-Wallis rank sum test

data:  cd45.model.4$cd45 by cd45.model.4$lobulardic
Kruskal-Wallis chi-squared = 6.5514, df = 1, p-value = 0.01048
```

Anexo 15 Test de Kruskal-Wallis para Inflamación Batts Ludwig Dicotomica y CD45.

```
        Kruskal-Wallis rank sum test

data:  cd45.model.5$cd45 by cd45.model.5$battsludwigdic
Kruskal-Wallis chi-squared = 22.523, df = 1, p-value = 2.077e-06
```

Anexo 16 Test de Kruskal-Wallis para Esteatosis y FPA.

```
        Kruskal-Wallis rank sum test

data:  fpa.model.1$fpa by fpa.model.1$esteatosis
Kruskal-Wallis chi-squared = 93.769, df = 3, p-value < 2.2e-16
```

Anexo 17 Comparación entre categorías mediante la prueba de Wilcoxon entre Esteatosis y FPA.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  fpa.model.1$fpa and fpa.model.1$esteatosis

  0       1       2
1 3.4e-07 -       -
2 3.9e-12 0.06099 -
3 7.5e-13 5.1e-07 0.00031

P value adjustment method: BH
```

Anexo 18 Test de Kruskal-Wallis para Esteatosis dicotomizada y FPA.

```
        Kruskal-Wallis rank sum test

data:  fpa.model.2$fpa by fpa.model.2$estdic
Kruskal-Wallis chi-squared = 58.344, df = 1, p-value = 2.2e-14
```

Anexo 19 Test de Kruskal-Wallis para Deugniers y Porcentaje de Hierro.

```
        Kruskal-Wallis rank sum test

data:  deug.model.1$deug by deug.model.1$deugniers
Kruskal-Wallis chi-squared = 131, df = 10, p-value < 2.2e-16
```

Anexo 20 Comparación entre categorías mediante la prueba de Wilcoxon entre Deugniers y el porcentaje de hierro.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  deug.model.1$hierrototal and deug.model.1$deugniers

   0       1       3       4       6       7       9       10      12      15
1  1.00000 -       -       -       -       -       -       -       -       -
3  0.48151 0.64818 -       -       -       -       -       -       -       -
4  0.00202 0.00595 0.05086 -       -       -       -       -       -       -
6  0.02998 0.15445 0.43137 0.94712 -       -       -       -       -       -
7  0.00044 0.00202 0.00450 0.54321 0.58757 -       -       -       -       -
9  0.00432 0.00946 0.02193 0.41026 0.45525 0.81345 -       -       -       -
10 0.50000 0.54321 0.69841 0.95652 1.00000 0.66667 0.60189 -       -       -
12 0.60189 0.69841 0.72368 0.81345 0.87302 0.93023 0.95652 1.00000 -       -
15 0.39348 0.43137 0.41026 0.50000 0.54321 0.50000 0.60189 1.00000 0.72368 -
18 0.39348 0.54321 0.53140 0.81345 0.72368 1.00000 1.00000 1.00000 1.00000 1.00000

P value adjustment method: BH
```

Anexo 21 Test de Kruskal-Wallis para Scheuer y Porcentaje de Hierro.

```
        Kruskal-Wallis rank sum test

data:  sche.model.2$sche by sche.model.2$scheuer
Kruskal-Wallis chi-squared = 131, df = 4, p-value < 2.2e-16
```

Anexo 22 Comparación entre categorías mediante la prueba de Wilcoxon entre Esteatosis y el porcentaje de hierro.

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  sche.model.2$hierrototal and sche.model.2$sche

  0       1       2       3
1 0.01124 -       -       -
2 0.03070 0.73556 -       -
3 0.00076 0.54635 0.35250 -
4 0.01124 0.35250 0.25486 0.87880

P value adjustment method: BH
```

Anexo 23 Test de Kruska-Wallis para Esteatosis dicotomica y Porcentaje de Hierro.

```
        Kruskal-Wallis rank sum test

data:  sche.model.3$scheuerdic by sche.model.3$hierrototal
Kruskal-Wallis chi-squared = 108, df = 108, p-value = 0.4819
```

Anexo 24 Matriz de confusión de Fibrosis Score. Predicción/Real (P/R).

P/R	0	1	2	3	4
0	21	6	8	2	0
1	3	5	1	1	0
2	3	5	1	1	0
3	0	2	6	7	4
4	1	0	0	2	0

Anexo 25 Matriz de confusión de Fibrosis Score Avanzada.

P/R	0	1
0	50	4
1	18	19

Anexo 26 Matriz de confusión de Inflamación Lobular.

P/R	0	1	2	3
0	5	9	4	0
1	16	23	11	3
2	3	9	7	3
3	0	0	0	0

Anexo 27 Matriz de confusión de la inflamación Lobular avanzada.

P/R	0	1
0	52	17
1	13	11

Anexo 28 Matriz de confusión de la inflamación Portal.

P/R	0	1
0	21	8
1	15	50

Anexo 29 Matriz de confusión de inflamación Batts Ludwig.

P/R	0	1	2	3	4
0	14	9	5	3	1
1	7	10	3	3	1
2	3	3	3	6	4
3	4	0	4	1	3
4	0	2	1	1	1

Anexo 30 Matriz de confusión de inflamación Batts Ludwig avanzada.

P/R	0	1
0	47	8
1	21	15

Anexo 31 Matriz de confusión de Esteatosis.

P/R	0	1	2	3
0	31	8	1	0
1	9	3	5	0
2	0	5	9	6
3	0	0	6	14

Anexo 32 Matriz de confusión de Esteatosis avanzada.

P/R	0	1
0	54	6
1	2	35

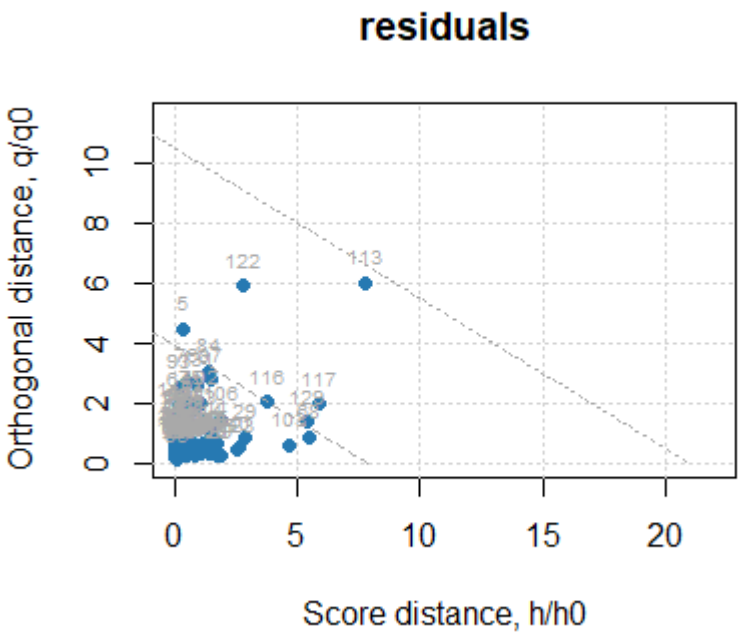
Anexo 33 Matriz de confusión de Scheuer.

P/R	0	1	2	3	4
0	44	3	1	3	0
1	4	1	1	0	0
2	6	0	2	0	1
3	3	0	1	0	0
4	7	2	1	2	2

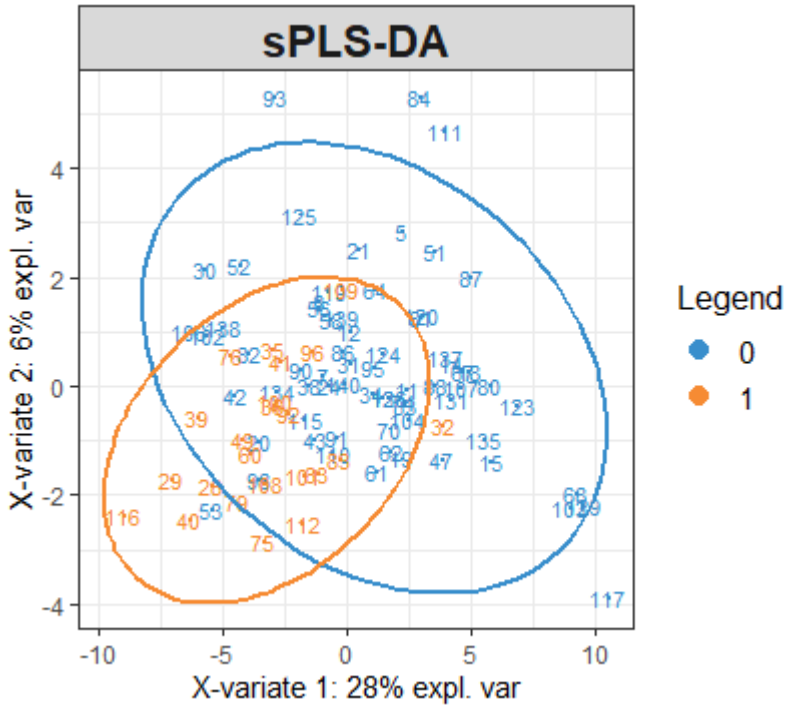
Anexo 34 Matriz de confusión de Scheuer avanzada.

P/R	0	1
0	23	17
1	17	27

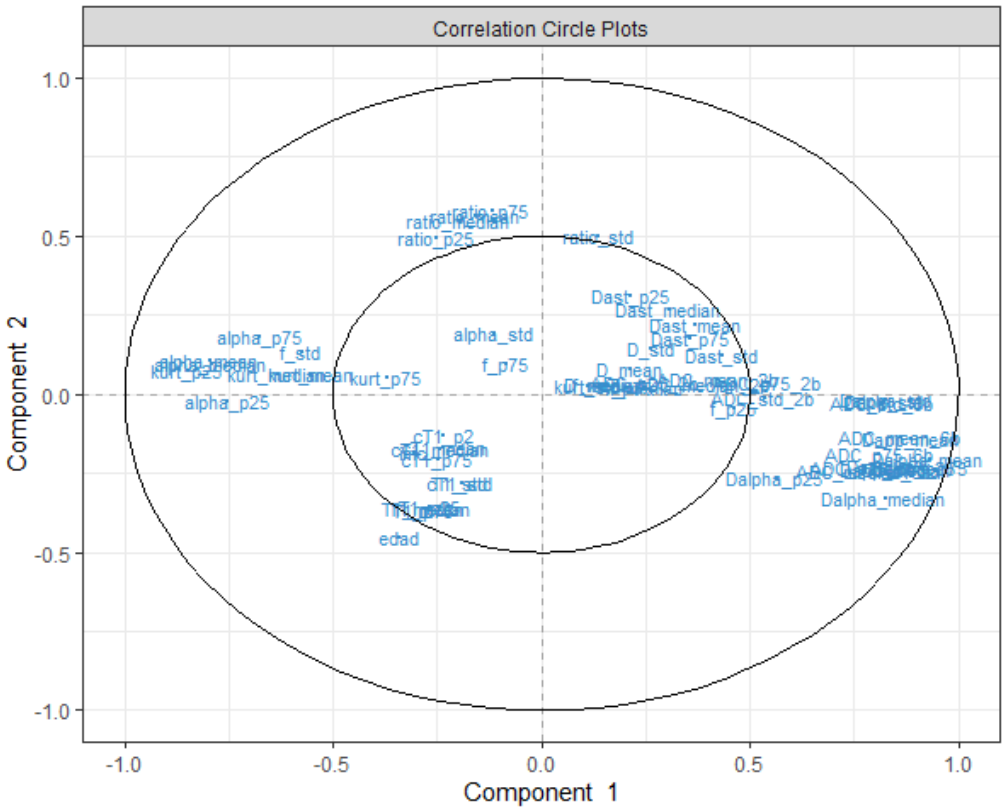
Anexo 35 Análisis de residuos del modelo sPLS-DA para la Fibrosis score avanzada.



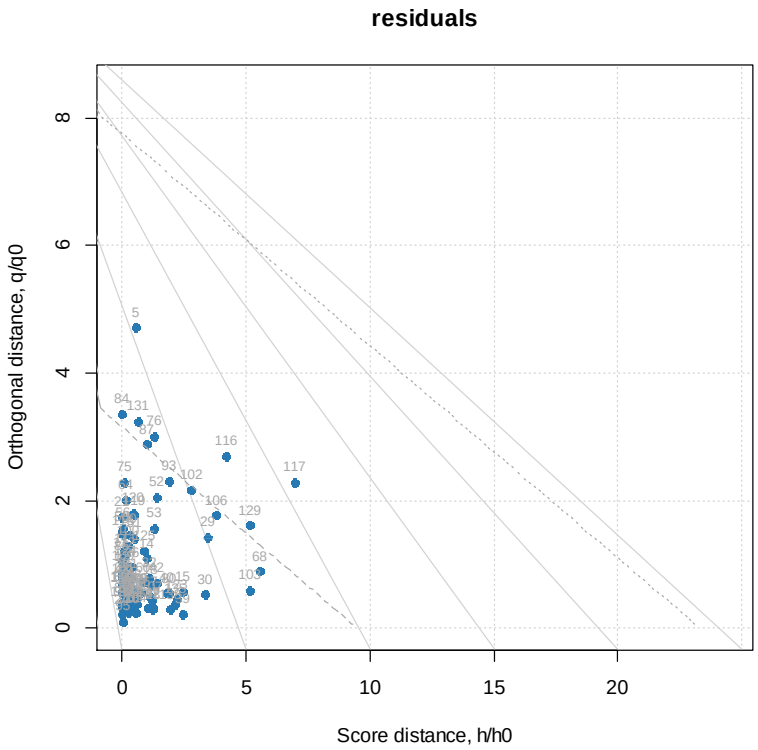
Anexo 36 Gráfico de Scores tras la eliminación de las observaciones con residuos elevados.



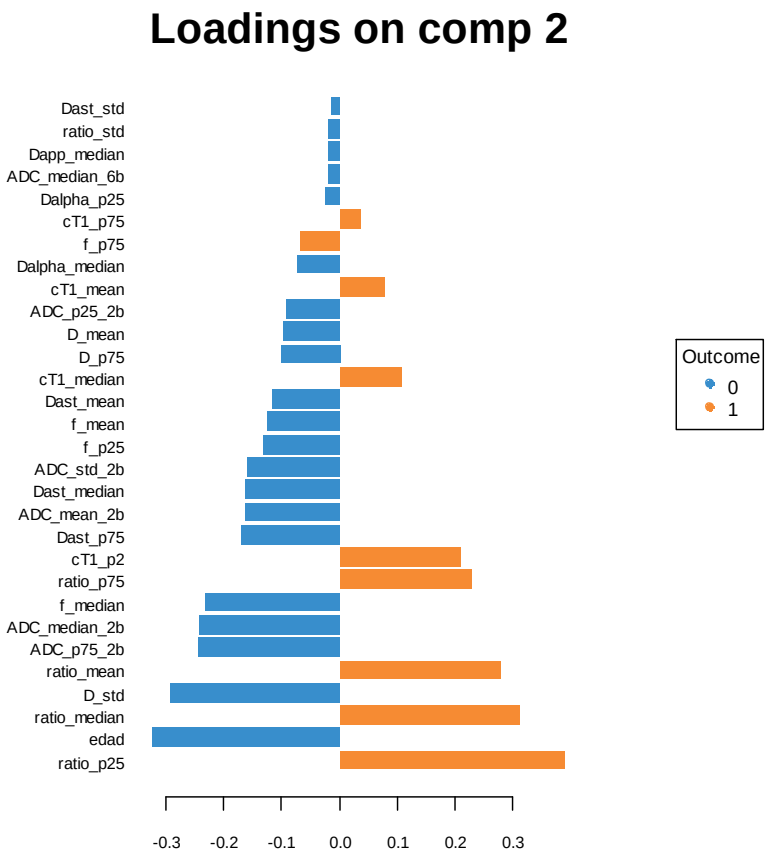
Anexo 37 Gráfico de loadings.



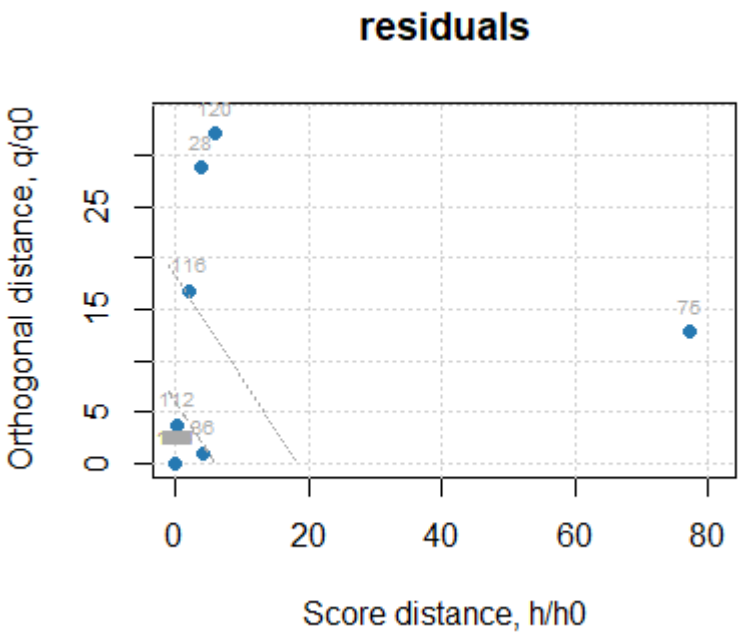
Anexo 38 Gráfico de Residuos SPE y T2 para la variable Batts Ludwig. Se han eliminado las observaciones 122 y 113.



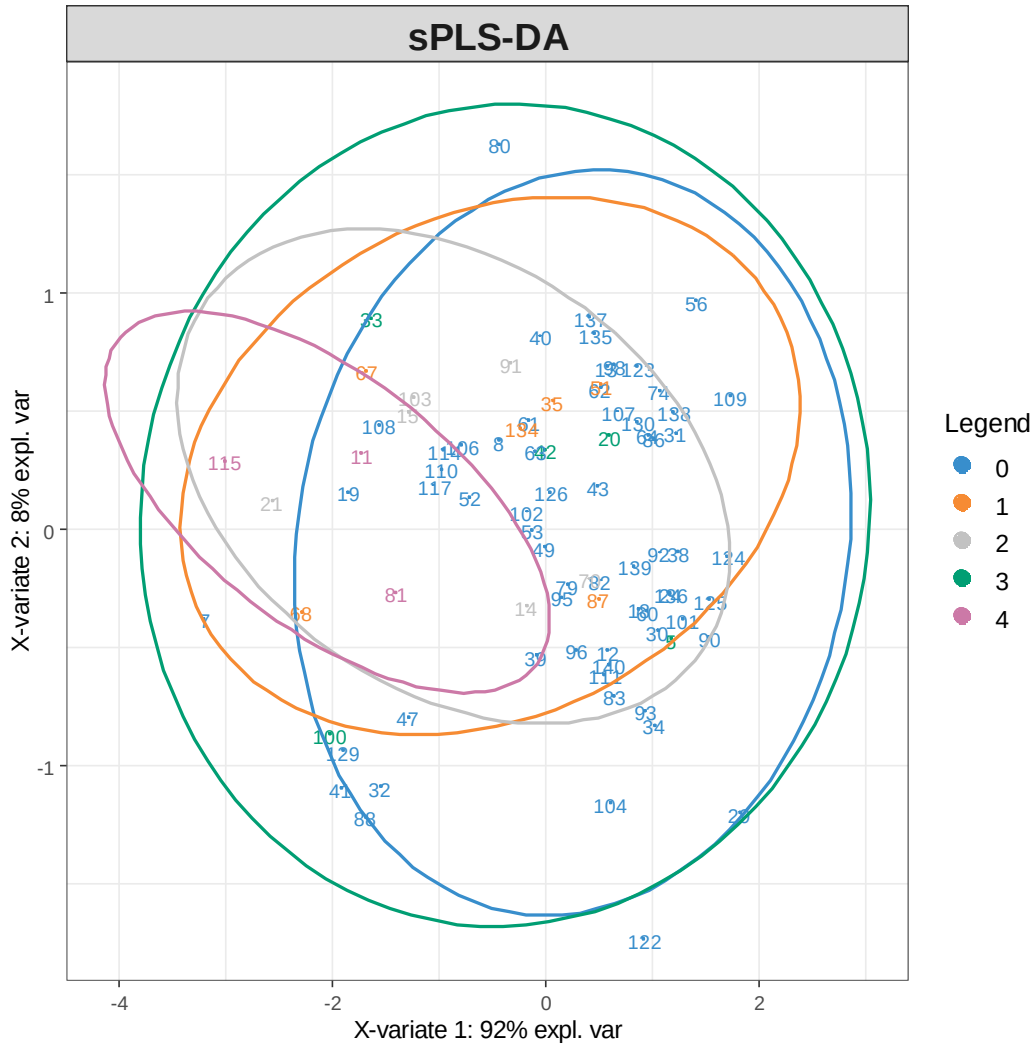
Anexo 39 Variables más influyentes en la segunda dirección principal.



Anexo 40 Gráfico de residuos SPE y T2 para la variable Scheuer.



Anexo 41 Gráfico de Scores de la variable Scheuer 0-4.



Anexo 42 Gráfico de scores de Scheuer dicotómica tras la eliminación de las observaciones extremas con residuos elevados.