UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria Agroalimentària
i de Biosistemes de Barcelona

Predicción del perfil lipoproteico en modelos animales

Trabajo de final de grado

Ingeniería de Sistemas Biológicos

Autor: Javier Barbazza Izquierdo
Tutor académico: Clara Prats Soler
Tutor profesional: Enrique Ozcariz Garcia
Co-tutor: Eduardo Dominguez Sala
Fecha de inicio: 14/02/2022
Fecha final: 07/07/2022

Resum

Les malalties cardiovasculars són, avui dia, molt prevalents en les poblacions arreu del món i s'espera que la tendència continuï en augment durant els anys que estan per venir. Per tant, és d'interès general desenvolupar nous mètodes que capacitin als professionals per assessorar amb fiabilitat en funció del perfil cardiovascular dels pacients. Gràcies a la metabolòmica s'han aconseguit noves tècniques analítiques com la ressonància magnètica nuclear (RMN) que s'han aplicat per predir lípids estàndard altament relacionats amb les malalties cardiovasculars. El present treball desenvolupa i avalua models de predicció PLS (Partial Least Squares regression) basats en la ressonància magnètica nuclear de protons amb la finalitat d'obtenir una quantificació de colesterol i triglicèrids totals en mostres de plasma animal de tres conjunts (porcs, N = 230 on N és el nombre de mostres). Primer es van efectuar estudis estadístics descriptius bàsics per comprendre millor el comportament de les mostres. Després es va efectuar un PCA (Principal Components Analysis) que indicaria com se separen les mostres segons les diferents variables. Després d'observar els resultats aconseguits fins al moment es va decidir excloure les mostres del conjunt 3, ja que no eren comparables a les mostres dels altres dos. A continuació es van realitzar gràfics STOCSY (Statistical TOtal Correlation SpectroscopY) per veure la correlació dels espectres de RMN i la variable de predicció a estudiar i per últim, a partir de totes les dades aconseguides, es va procedir a crear els models PLS. Es van utilitzar dos mètodes, el primer utilitzava el 70% de les dades per a l'entrenament i el 30% restant per a la prova de validació, el segon consistia a seleccionar el 10% dels valors de les variables més alts, el 10% dels valors més baixos, el 50% dels valors al atzar per a l'entrenament i el 30% restant, igual que per al primer mètode, es va utilitzar per a la prova de validació. Es va aplicar a les dades dos tipus de preprocessats per tal d'observar amb quin s'assolien millors resultats. Aplicant el preprocessat Autoscale es van aconseguir resultats prometedors amb models que arribaven a un ajust de predicció del 0,97 (R=0,97) per al colesterol pel que fa als triglicèrids es van observar millors resultats fent servir el preprocessat Mean centering el qual va donar un ajust màxim de 0,86 (R=0,86). Els resultats del treball mostren que és possible emprar els models de predicció PLS per a quantificar lipoproteïnes en mostres de plasma de porc de forma fiable.

Resumen

Las enfermedades cardiovasculares son, hoy en día, muy prevalentes en las poblaciones de todo el mundo y se espera que la tendencia siga en aumento en los años venideros. Por lo tanto, es de interés general desarrollar nuevos métodos que capaciten a los profesionales para asesorar fiablemente en función del perfil cardiovascular de los pacientes. Gracias a la metabolómica se han conseguido nuevas técnicas analíticas como la resonancia magnética nuclear (RMN) que se han aplicado para predecir lípidos estándares altamente relacionados con las enfermedades cardiovasculares. El presente trabajo evalúa y desarrolla modelos de predicción PLS (Regresión de mínimos cuadrados parciales) basados en la resonancia magnética nuclear de protones con la finalidad de obtener una cuantificación del colesterol y los triglicéridos totales en muestras de plasma animal de tres conjuntos (cerdos, N = 230 donde N es el número de muestras). Primero se efectuaron estudios estadísticos descriptivos básicos para comprender mejor el comportamiento de las muestras. Después se efectuó un PCA (Análisis de componentes principales) que indicaría cómo se separan las muestras según diferentes variables. Tras observar los resultados obtenidos hasta el momento se decidió excluir las muestras del conjunto 3 ya que estos no eran comparables a las muestras de los otros dos. A continuación se realizaron gráficos STOCSY (Espectroscopia estadística de correlación total) para observar la correlación de los espectros de RMN y la variable de predicción a estudiar y por último, a partir de todos los datos obtenidos, se procedió a crear los modelos de predicción PLS. Se usaron dos métodos, el primero usaba el 70% de los datos para el entrenamiento y el 30 % restante para la prueba de validación, el segundo consistia en seleccionar el 10% de los valores de las variables más elevados, el 10% de valores más bajos, el 50% de los valores aleatorios para el entrenamiento y el 30% restante, al igual que en el primer método, se usó para la prueba de validación. Se aplicó dos tipos de preprocesados a los datos para observar con cual se obtenían mejores resultados. Aplicando el preprocesado Autoscale se consiguieron resultados prometedores con modelos que llegaban a un coeficiente de correlación de 0,97 (R=0,97) para colesterol, en cuanto a los triglicéridos se observaron mejores resultados usando el preprocesado Mean centering que dió un coeficiente de correlación máximo de 0,86 (R=0,86). Los resultados del trabajo muestran que es posible usar los modelos de predicción PLS para cuantificar lipoproteínas en muestras de plasma de cerdo fiablemente.

Abstract

Nowadays cardiovascular diseases have a high influence on the world's population and this trend is expected to increase over the years to come. Therefore, there is a general interest to develop new methods that enable professionals to advise reliably depending on the patient's cardiovascular profile. Thanks to metabolomics, new techniques have been achieved such as the nuclear magnetic resonance (NMR) which has been applied to predict standar lipids highly related with cardiovascular diseases. The present project evaluates and develops PLS (Partial Least Square) prediction models based on nuclear magnetic resonance of protons with the aim to achieve a quantification of the total cholesterol and triglycerides of three sets of animal plasma samples (pigs, *N* = 230 where *N* is the number of samples). First basic descriptive statistics were performed in order to better understand the behavior of the samples. Afterwards, a PCA (Principal Components Analysis) was performed showing how the samples spread according to different variables. Then, it was decided to exclude the samples belonging to the set number 3 since they were not comparable to the samples of the other two sets. Next STOCSY (Statistical TOtal Correlation SpectroscopY) plots were created to observe the correlation between the NMR spectra and the studied prediction variable, lastly, from all the obtained data, the PLS predictions models were created. For this purpose, two methods were used, the first one used 70% of the values as a training set and the remaining 30% as a testing set, the second one consisted on selecting the 10% of the highest values, the 10% of the lowest values, 50% of the values which were randomly selected as a training set and the remaining 30% were used as a testing set. Two types of preprocessing were applied to the data in order to identify which was the one returning the best results, when applied, the Autoscale preprocessing got promising results achieving models with a 0.97 (R=0,97) of goodness of fit to cholesterol. Regarding triglycerides, better results were given when the Mean centering preprocessing was applied giving a maximum fitness of 0.86 (R=0.86). After the completion of this project, the creation of PLS prediction models to quantify lipoproteins in pig plasma samples reliably is clearly possible.

Sumario

Índice de figuras

Índice de tablas

Símbolos y acrónimos

Acrónimos:

1D: Una dimensión

2D: Dos dimensiones

1H-RMN: Resonancia Magnética Nuclear de protones en una dimensión (1H NMR en inglés)

Apo: Apolipoproteína

C: Colesterol

CT o TC: Colesterol total o Total cholesterol en inglés

Desv. Est.: Desviación Estándar

DOSY-NMR: Espectroscopía de resonancia magnética nuclear por orden de difusión o Diffusion Ordered Nuclear Magnetic Resonance Spectroscopy en inglés

HDL: Lipoproteína de alta densidad o High Density Lipoprotein en inglés

H-RMN: Resonancia Magnética Nuclear de protones (H NMR en inglés)

IDL: Lipoproteína de densidad intermedia o Intermediate Density Lipoprotein en inglés

JLR: Journal of Lipid Research

LDL: Lipoproteína de baja densidad o Low Density Lipoprotein en inglés

LED: Longitudinal Eddy-Current Delay

MS: Espectroscopía de masas o Mass Spectroscopy en inglés

NOESY: Nuclear Overhauser Effect SpectroscopY

PC: Componente principal o Principal Component en inglés

PCA: Análisis de componentes principales o Principal Components Analysis en inglés

PLS: Regresión de Mínimos Cuadrados Parciales o Partial Least Squares en inglés

Q1, Q2 y Q3: Cuartil 1, Cuartil 2 y Cuartil 3 o Quartile 1, 2 and 3 en inglés

R: Coeficiente de correlación de Pearson o Pearson correlation coefficient en inglés

RMN o NMR: Resonancia Magnética Nuclear (NMR en inglés)

STOCSY: Espectroscopia estadística de correlación total o Statistical total correlation spectroscopy en inglés

TG: Triglicéridos totales

VLDL: Lipoproteína de muy baja densidad o Very Low Density Lipoprotein en inglés

VAP: Auto-perfilado vertical o Vertical Auto Profile en inglés

Trabajo de final de grado en asociación con la empresa Biosfer Teslab S.L. como estudiante en prácticas bajo la supervisión de Enrique Ozcariz Garcia

Agradecimientos

Me gustaría agradecer al equipo de Biosfer Teslab por su acogida a la empresa como estudiante en prácticas y por darme la oportunidad de vivir esta experiencia. Todo el equipo ha mostrado mucho apoyo y se han ofrecido a ayudar cuando encontraba alguna dificultad. En particular me gustaría dar las gracias a mis tutores de la empresa, Enrique Ozcariz Garcia, Daniel Rodriguez Romeu y Sara Samino Gené por dedicar su tiempo a enseñarme nuevos conceptos y guiarme a lo largo de este proyecto desde el primer momento.

Querría agradecer también la labor de Eduardo Domínguez Sala que me ha ayudado enormemente, como tutor, en el desarrollo del trabajo escrito y asesorado con puntos de vista diferentes que ignoraba y han dotado a este trabajo de fluidez y una nueva esencia.

Otra persona a la que me gustaría agradecer es a Clara Prats Soler por aceptar ser mi tutora académica y aportar sugerencias de gran utilidad en la estructura del trabajo para que los lectores puedan seguirlo con facilidad.

Por último, agradecer a todos mis familiares y amigos que me han apoyado durante el desarrollo del trabajo de fin de grado y que siempre me han animado a seguir adelante.

1. Introducción

1.1. Situación actual

Según datos oficiales de la Organización Mundial de la Salud (OMS), actualmente, la principal causa de muerte en el mundo son las cardiopatías. Un estudio publicado en 2019 por dicha organización mostró que el 32% de muertes a nivel mundial, aproximadamente 17,9 millones de personas, son a causa de las cardiopatías **[1]**.

Tales datos hacen visible la necesidad de un buen sistema sanitario que promueva hábitos de vida saludable empleando estrategias de medicina preventiva y que dote a su vez a los centros sanitarios de los recursos necesarios para la detección precoz y el seguimiento adecuado de dichas patologías.

Una de las causas más relevantes para el desarrollo de enfermedades cardiovasculares es la alteración de niveles en el plasma sanguíneo de ciertas moléculas relacionadas con el metabolismo energético. Entre estas, se puede destacar los carbohidratos, como por ejemplo la glucosa, cuya desregularización supone la aparición de diabetes, o ciertos lípidos como el colesterol y los triglicéridos. Aunque el cuerpo humano posee las herramientas bioquímicas necesarias para la producción de este tipo de biomoléculas, frecuentemente, la alteración de los niveles en plasma de estas suele estar asociada con hábitos de vida insalubres tales como una dieta hipercalórica o el sedentarismo.

1.2. Lipoproteínas

Tanto colesterol como triglicéridos, debido a su naturaleza química, son moléculas altamente hidrofóbicas, por lo que su transporte en plasma, que es un medio altamente hidrofílico por su alto contenido en agua, no ocurre en su forma libre, sino que son transportadas en forma de complejos macromoleculares conocidos como lipoproteínas plasmáticas. Una lipoproteína es una partícula compleja formada por el empaquetamiento de distintas moléculas. En la **Figura 1-1** se presenta un esquema representativo de una lipoproteína. Las lipoproteínas presentan una monocapa externa de fosfolípidos, los cuales orientan el grupo fosfato (parte polar) al exterior, en contacto con el plasma, y sus colas apolares, formadas por ácidos grasos, hacia el interior. En esta monocapa de fosfolípidos encontramos además apolipoproteínas y colesterol libre **[2,3]**. En el interior de una partícula lipoproteica encontramos acumulados ésteres de colesterol (producto de la unión de un ácido graso

y una molécula de colesterol) y triglicéridos. Por tanto, las lipoproteínas son las encargadas de transportar los triglicéridos, colesterol y fosfolípidos, entre otros tipos de lípidos, a través del plasma sanguíneo. Con la finalidad de poder predecir, prevenir y tratar las enfermedades cardiovasculares es imprescindible entender qué son las lipoproteínas y qué función tienen.

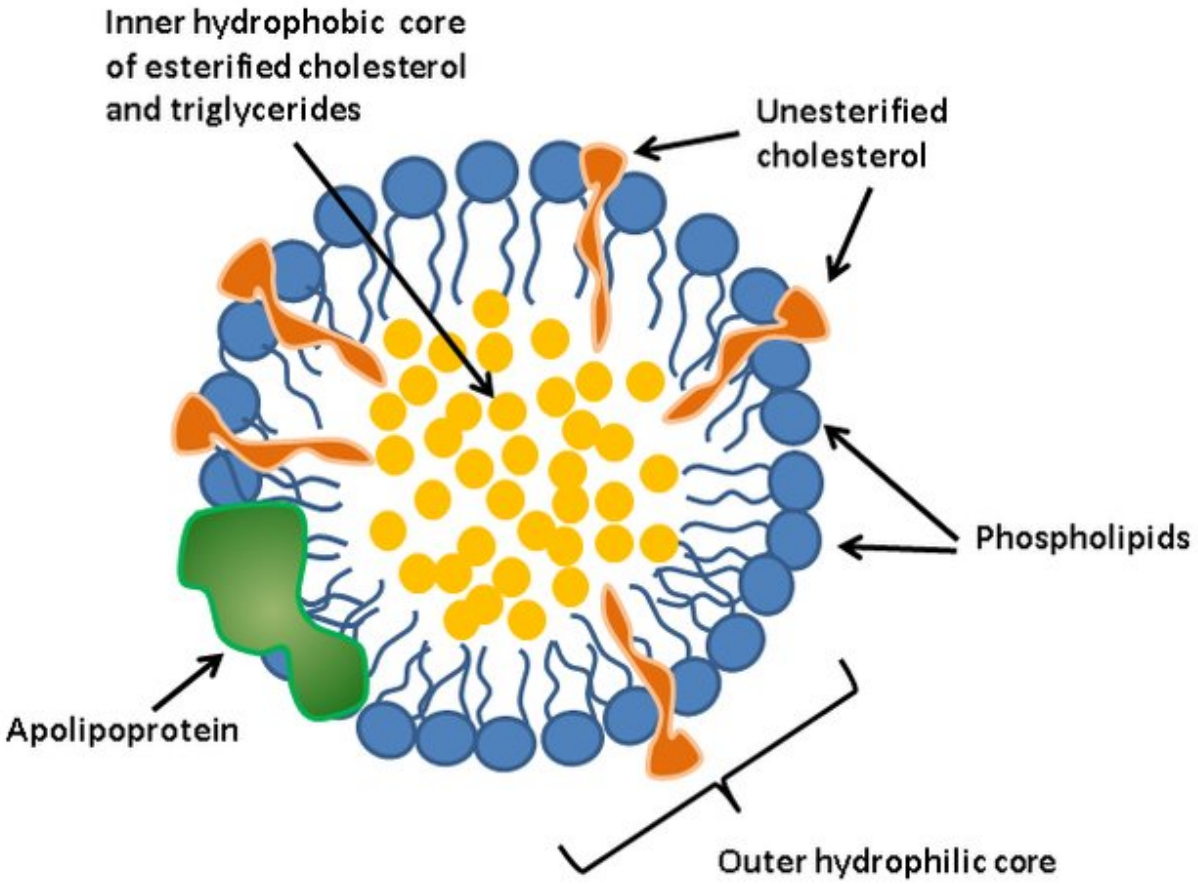


Figura 1-1. Esquema de una lipoproteína **[4]**.

Aunque algunos autores dividen las lipoproteínas hasta en siete clases **[3]**, cuyas diferencias radican esencialmente en el tamaño, densidad y cantidad de moléculas lipídicas contenidas, en este proyecto solo nos centraremos en los cuatro tipos con los que trabaja la empresa Biosfer Teslab. Estos son las VLDL (very low density lipoprotein/ lipoproteína de muy baja densidad), IDL (intermediate density lipoprotein/lipoproteína de densidad intermedia), LDL (low density lipoprotein/ lipoproteína de baja densidad y HDL (high density lipoprotein/ lipoproteína de alta densidad).

Es importante saber tanto la composición como la función que presentan las diferentes clases de las que vamos a hablar en el estudio. Las partículas VLDL son producidas en el hígado, tienen una composición rica en triglicéridos que son transportados por todo el organismo. Cuando los triglicéridos son extraídos de estas partículas por los músculos y el tejido adiposo se forman las

partículas IDL que son ricas en colesterol. La siguiente clase son las LDL que derivan de las VLDL y las IDL, las cuales están muy enriquecidas con colesterol y son las mayores transportadoras del mismo. Son las que comúnmente se conocen como “colesterol malo” debido a su prolongado tiempo de retención en la circulación además que por su tamaño son capaces de entrar fácilmente en las paredes arteriales y obstruir la circulación. Una de sus funciones está relacionada con la aterogenia que es la respuesta fisiológica a la rotura de un vaso sanguíneo y que resulta en la inflamación, la reparación y el engruese de dichos vasos **[3]**. Otra característica a tener en cuenta es que estas partículas son susceptibles a la oxidación, factor que puede aumentar la acumulación de colesterol en el organismo. Por último, las partículas HDL, también conocidas como “colesterol bueno”, son las encargadas de revertir el transporte de colesterol en los tejidos periféricos hacía el hígado reduciendo así la aterogenia causada por las partículas LDL y IDL. También tiene propiedades antioxidantes además de ser ricas en colesterol y en fosfolípidos. Son consideradas como muy heterogéneas. Así pues, la proporción de estos distintos tipos de lipoproteínas presente en el plasma sanguíneo son indicadores del riesgo de padecer enfermedades cardiovasculares.

Un componente esencial de las lipoproteínas son las apolipoproteínas. Las apolipoproteínas son proteínas que poseen distintas funciones esenciales para las lipoproteínas como por ejemplo mantener la cohesión de estas partículas **[5]**, guiar la formación de estas o actuar como señal para su reconocimiento y transporte al interior celular desde el plasma, actuar como coactivadores o inhibidores de enzimas como la LPL (por sus siglas en inglés LipoProtein Lipase) o la lipasa hepática **[6, 7]** que cataliza la hidrólisis de quilomicrones y triglicéridos VLDL. Como se ha mencionado anteriormente, están situadas en la membrana de las lipoproteínas. Hay seis clases destacables: Apo (apolipoproteína) A, Apo B, Apo C, Apo D, Apo E y Apo H, estás a su vez tienen varias subclases de las cuales no se hablará en este trabajo. Las apolipoproteínas A componen la mayor parte de las lipoproteínas de alta densidad, puede actuar como biomarcador de estas lipoproteínas y son capaces de modificar los niveles de triglicéridos en el corazón entre otras funciones. Las Apo B son las principales componentes de las lipoproteínas de baja densidad, altos niveles de esta clase de proteína indican riesgos para el corazón que se pueden estimar gracias al cálculo de la relación entre los niveles de Apo B respecto a los de Apo A. Las Apo C se encargan de activar o inhibir la lipasa cuya presencia ayuda a la absorción de grasa descomponiendola en ácidos grasos, un déficit de esta proteína puede conllevar varios problemas cardiovasculares por acumulación de triglicéridos. Las apolipoproteínas D son componentes de las lipoproteínas de alta densidad con propiedades antioxidantes y antiinflamatorias. Las Apo E están relacionadas con el reconocimiento de las lipoproteínas de densidad intermedia situadas en el hígado, pero también es muy importante para

evitar enfermedades cerebrales; al igual que las Apo A son capaces de modificar los niveles de triglicéridos. Las apolipoproteínas H están relacionadas con la coagulación de la sangre y la producción de algunos anticuerpos, también se las conoce como glicoproteínas I.

1.3. Modelos de experimentación animal

El uso de modelos animales en investigación es una práctica habitual para la comprensión y la prevención de enfermedades **[8]**. Especies de modelos animales de tamaño pequeño como el ratón (*Mus musculus*), se utilizan para conseguir este objetivo. Desafortunadamente, trabajar con estos modelos que en algunos aspectos difieren tanto de las condiciones fisiológicas humanas, hace que la eficacia en ensayos clínicos se vea reducida drásticamente o requieran un proceso de análisis experimental más exhaustivo. El interés en el uso de ratones para dichos estudios reside en distintos aspectos. En primer lugar, su bajo coste económico respecto a especies de mayor tamaño ya que requieren menos espacio e instalaciones más simples. Otro factor de interés es su alta capacidad reproductiva, la cual permite obtener varias generaciones a lo largo de un año natural. Por otro lado, las características previamente mencionadas han permitido realizar numerosas investigaciones con este tipo de animales modelo, obteniendo así una considerable cantidad de información genética sobre ellos, hecho que ha permitido la creación de múltiples variedades modificadas genéticamente con relativa facilidad. Estas variedades modificadas genéticamente han sido base fundamental en el estudio de diversas patologías de origen genético, incluidas las cardiovasculares. Por último, también hay que tener en cuenta que éticamente experimentar con animales más alejados filogenéticamente de la especie humana está más aceptado que con modelos de mamíferos más cercanos a esta.

A pesar de sus numerosas ventajas **[9]**, en ciertas ocasiones, el uso de ratones para investigar enfermedades humanas se debe desestimar debido a las diferencias fisiológicas como se ha mencionado previamente. Un modelo animal mucho más cercano para investigar la respuesta fisiológica en humanos es el cerdo (*Sus scrofa domesticus*), ya que posee un sistema cardiovascular con características mucho más similares a las del ser humano respecto otros modelos. Algunas de estas características son: posee un corazón similar al humano, son más tolerantes a técnicas invasivas y se les puede tratar con dosis similares a las que se usaría en humanos, por el contrario, aquellas ventajas que teníamos con los ratones ya no se aplican.

1.4. Estado del arte

El trabajo con animales de experimentación contribuye al estudio y desarrollo de estrategias de análisis de lipoproteínas que posteriormente serán aplicadas a pacientes. Por lo tanto, el uso de modelos animales de experimentación es fundamental. De hecho, existen diversas investigaciones previas empleando modelos animales las cuales usan diversas técnicas para llegar al mismo objetivo.

Por ejemplo, un estudio publicado en JLR (Journal of Lipid Research) el 2012 por Wu Yin y su equipo trataba de usar modelos animales para predecir la dislipidemia **[10]**, el desequilibrio de lípidos como por ejemplo el colesterol de baja densidad (LDL-C) **[11]**, mediante muestras de plasma. Las muestras eran un total de 24, 5 de ratones, 6 de especies no primates y 4 especies de primates no humanos y las 9 restantes de humanos. En el experimento se usó la cromatografía para separar las lipoproteínas, posteriormente midieron el colesterol total con un detector de colesterol seguido de una detección espectrométrica de los reactivos a partir de los cuales se consiguieron sacar los niveles de VLDL, LDL-C (colesterol) y HDL-C multiplicandolos por el área relacionada con los picos de estas partículas a partir de la cromatografía y luego se dividió el área de todos los picos por el colesterol total.

Otro estudio publicado en 2007 por Wendy Mercedes Rauw **[12]**, miembro del Consejo Superior de Investigaciones Científicas (CSIC) en España, investigó la relación, en cerdos, entre la ingesta de comida y el colesterol en el cuerpo, en particular aquellos casos en que la ingesta está relacionada con la obesidad y cómo los patrones alimenticios están relacionados con la calidad de la comida. Para determinar las concentraciones totales de colesterol y triglicéridos: VLDL, LDL, IDL, HDL, se hizo uso del programa informático Technicon Chem 1. Para medir los niveles de HDL se usó el sobrenadante obtenido en la precipitación de aquellas lipoproteínas que contenían Apo B mientras que para conseguir los niveles de LDL se empleó la ecuación de Friedewald **[13]**.

Por último, en un estudio publicado en 2007 por Hanne C. Bertram, una investigadora en la universidad de AARHUS en Dinamarca **[14, 15]**, se utiliza la caracterización avanzada de lipoproteínas mediante el uso de resonancia magnética nuclear (RMN) aunque se basa en una aproximación que es ligeramente diferente a la de este trabajo. Por lo tanto, el estudio del perfil lipídico y, sobre todo, del perfil avanzado de lipoproteínas en modelos animales es de gran interés en la comunidad científica, aunque la caracterización avanzada de lipoproteínas puede conllevar tanto a retos tecnológicos como retos biológicos debido a la dificultad en la estandarización de los métodos.

El proyecto de Biosfer Teslab contiene las mismas técnicas analíticas que el estudio de Hanne C. Bertram, pero difiere en el algoritmo que proporciona la caracterización avanzada. Biosfer los obtiene a partir de su innovador test Liposcale® que se basa en la deconvolución, mientras que el estudio de la universidad de AARHUS utiliza un algoritmo de predicción basado en regresiones lineales.

1.5. Métodos para la determinación de lipoproteínas

En este apartado se darán a conocer algunos métodos para determinar lipoproteínas y se verá en más profundidad la técnica analítica utilizada en el presente trabajo.

1.5.1. VAP

Existen diversas técnicas para estudiar las lipoproteínas cuantitativamente. Una de ellas es el método VAP (Vertical Auto Profile method). Esta técnica permite el análisis continuo de las enzimas del colesterol gracias a un espín con ultracentrifugado del gradiente de densidad, pero no es una técnica muy usada debido a que a la hora de observar los resultados se sobreponen por lo que perdemos resolución y no se consigue una correcta cuantificación de las lipoproteínas **[16]**.

1.5.2. Espectroscopia de masas

La espectroscopia de masas o *mass spectroscopy* (MS) **[17]**, es un método que se usa para determinar el peso molecular de las muestras permitiendo la identificación de compuestos desconocidos, cuantificar los compuestos que sí se conocen y determinar las estructuras y propiedades químicas de las moléculas.

Las moléculas con las que trabaja esta técnica son convertidas a iones en fase gaseosa y gracias a una fuente de ionización que mediante campos magnéticos y eléctricos externos posibilitan mover y manipular dichos iones. Una vez ionizados, los iones son clasificados según su relación masa-carga usando un analizador de masas. Por último, los iones previamente separados son medidos y enviados a un ordenador junto con los datos de la relación masa-carga y su abundancia relativa. Para obtener los resultados se crea un gráfico donde el eje *x* constituye la relación masa-carga y el eje *y* la abundancia relativa, se muestra un ejemplo en la **Figura 1-2.**

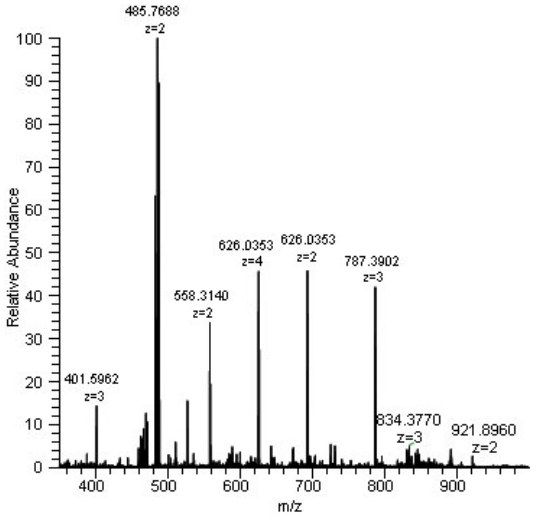


Figura 1-2. Ejemplo de un espectro de masa **[17]**

1.5.3. Resonancia Magnética Nuclear de Protones (H RMN)

Uno de los métodos más habituales y que se ha mencionado es la resonancia magnética nuclear de protones (H RMN). Esta técnica es de las más conocidas, se usa en infinidad de campos, pero muchos investigadores la usan para cuantificar las lipoproteínas en las muestras de plasma debido a que la resonancia separa frecuencias a razón del tamaño de las partículas situadas en los metilos de las partes centrales de la lipoproteína. Este método causa interés en la comunidad científica y se resiste a ser sustituido debido a que es una técnica no destructiva, ya que el tratamiento de las muestras es simple lo cual facilita su cuantificación. Otra ventaja que ofrece esta técnica es que permite la identificación de nuevos compuestos **[18]**. Con la finalidad de hacer menos abstracto el concepto de ^{1}H-RMN se ha adjuntado la **Figura 1-3** como ejemplo.

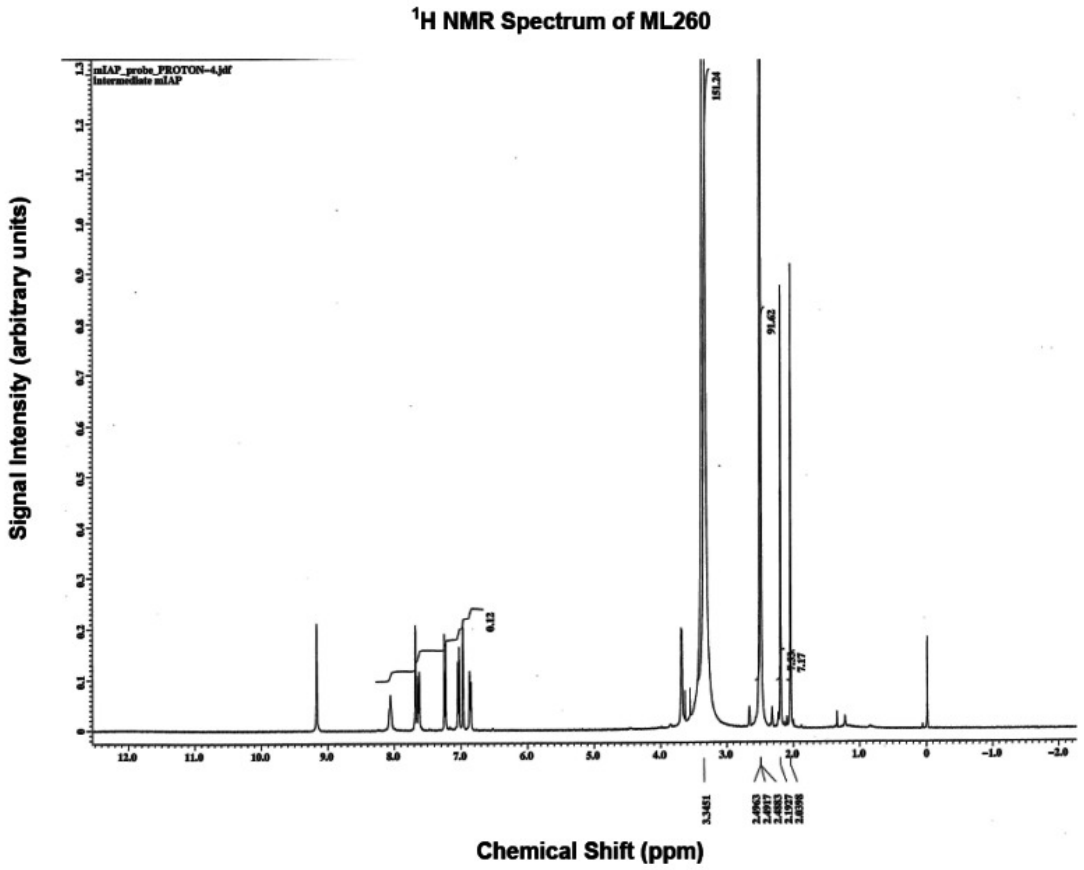


Figura 1-3. Ejemplo de espectro de ^{1}H-NMR de ML260 **[19]**

Dada su alta reproducibilidad y a la posibilidad de automatizar el proceso este método resulta altamente atractivo para proyectos a gran escala respecto de otras técnicas, habitualmente, más usadas en metabolómica, ya que, como particularidad, la espectroscopia NMR es capaz de detectar y caracterizar azúcares, ácidos orgánicos, alcoholes y otros componentes polares. Además, este recurso no se limita a biofluidos o extracciones de tejidos, sino que es posible estudiar muestras de tejidos intactos, sólidos, semisólidos, incluso es posible aplicar este método para diferentes elementos como C (Carbono), N (Nitrógeno), P (Fósforo) o H (Hidrógeno) siendo el último de particular interés para este proyecto. Cuando los núcleos de estos elementos interaccionan con el campo magnético, tienen la capacidad de girar adquiriendo dos orientaciones diferentes **[20]**, una que corresponde con el nivel más bajo de energía en relación al campo magnético y otro correspondiendo al más alto, cuando los núcleos son irradiados por un pulso de radiofrecuencia perpendicular al campo magnético provocando cambios en la orientación de los espines en los núcleos. Al finalizar la irradiación los núcleos vuelven a su estado original. Esta frecuencia de giro generada por el pulso de radiofrecuencia es captada y se usa posteriormente para proyectar un espectro de frecuencia e intensidad, donde la intensidad es proporcional al número de átomos, dicho proceso está representado en la **Figura 1-4.**

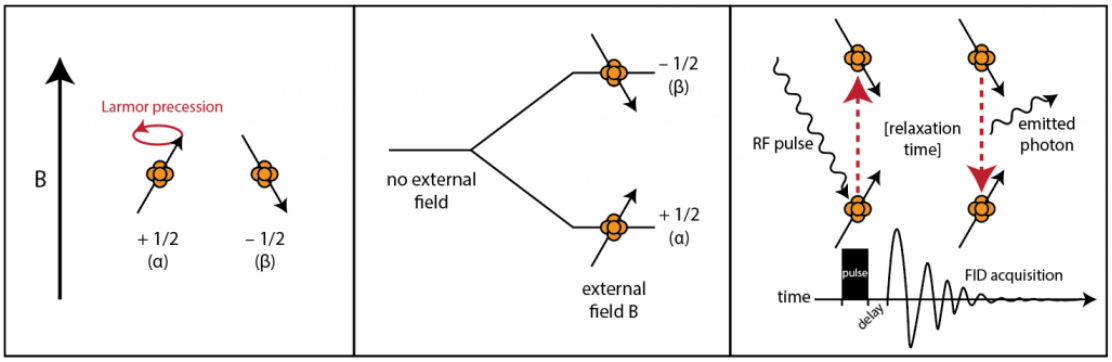


Figura 1-4. Funcionamiento de la resonancia magnética nuclear (RMN). **[21]**

La espectroscopia de resonancia magnética nuclear de protones o H-NMR, por sus siglas en inglés, es la más usada de las resonancias magnéticas nucleares dado que habitualmente los átomos de hidrógeno están presentes en los compuestos orgánicos. Por ende, los lípidos y en especial los triglicéridos y el colesterol (**Figura 1-5**), tienen una gran cantidad de átomos de este elemento en su estructura **[18, 20]**. Por ese motivo, esta técnica es de gran interés para el presente proyecto, ya que Biosfer Teslab trabaja con metabolómica, que es una disciplina que esta técnica permite estudiar, para poder predecir el perfil lipoprotéico de las distintas muestras. En esencia la metabolómica consiste en el estudio de los metabolitos. Otro rasgo característico de esta técnica es que está estrechamente ligada al uso de la tecnología y la informática ya que permite una mejor recopilación de datos para su posterior análisis e interpretación.

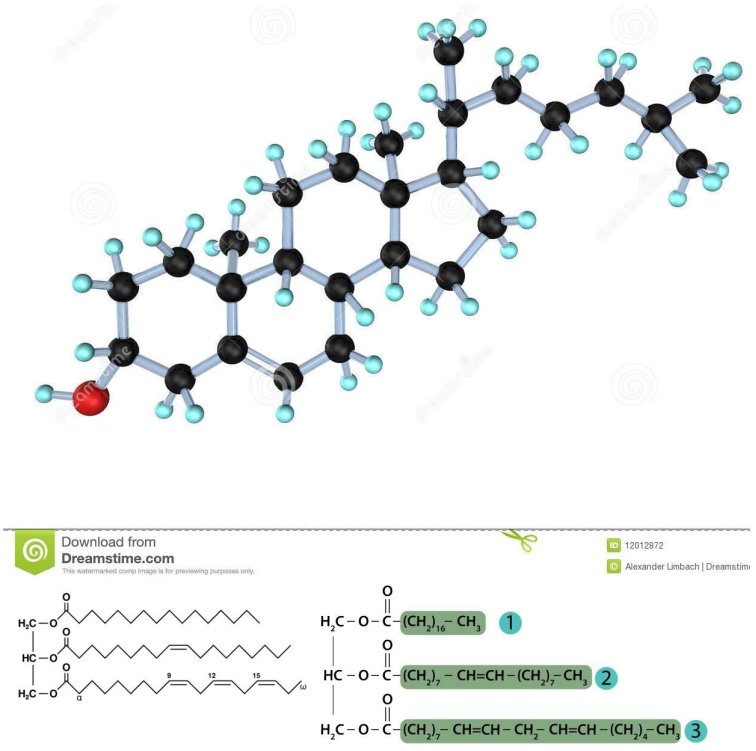


Figura 1-5. Estructura molecular del colesterol (arriba), las bolas azules representan los átomos de hidrógeno, las bolas negras los átomos de carbono y la roja el átomo de oxígeno. Esquemas moleculares de los triglicéridos (abajo) **[22,23,24]**.

Caracterización de lipoproteínas por RMN en una dimensión (1D)

Aunque la determinación de lipoproteínas por RMN se lleva haciendo hace ya algunos años, tanto a nivel de investigación como mediante pruebas comerciales, la mayoría de aproximaciones se limitan a los picos lipídicos y al análisis de lipoproteínas basados en resultados de modelos empíricos desarrollados mediante correlación entre el espectro crudo de RMN y medidas bioquímicas de laboratorio.

El análisis de lipoproteínas mediante la espectroscopia de RMN se basa en la siguiente propiedad física: dependiendo del tamaño de la partícula, los grupos metilo de los lípidos que viajan dentro de las lipoproteínas resuenan a frecuencias ligeramente diferentes en función del tamaño de la lipoproteína que los transporta, partículas más pequeñas resonando a frecuencias más bajas. Por lo tanto, las lipoproteínas pueden ser cuantificadas ya sea por descomposición de la señal de RMN del grupo metilo de los lípidos en señales individuales o utilizando métodos estadísticos sobre la totalidad de la envolvente de RMN para estimar las concentraciones lipídicas, que son directamente proporcionales a la intensidad de la señal. El primer método, descrito por Jeyarajah, Cromwell y Otvos [25], proporciona la concentración de partículas de las clases principales de lipoproteínas (es decir, VLDL, LDL, y HDL) y el número de partículas de nueve subclases, a partir de la estimación del tamaño de manera indirecta. Este método se basa en una biblioteca de espectros de RMN en 1D de clases de lipoproteínas previamente aisladas y en un algoritmo que ajusta su señal de RMN con los de muestras de suero o plasma.

Caracterización de lipoproteínas por RMN en dos dimensiones (2D)

Como alternativa a los métodos actuales de RMN basados en espectros 1D, Liposcale® aparece como un nuevo método para la caracterización de las lipoproteínas basado en espectroscopia de RMN de difusión en 2D **[26]** del suero o plasma sanguíneo.

La aproximación que utiliza Liposcale® es novedosa porque mediante la utilización de experimentos de RMN en 2D cuya señal está modulada por la difusión de las partículas en la mezcla (Diffusion Ordered Nuclear Magnetic Resonance Spectroscopy, DOSY-NMR) se pueden conocer las características hidrodinámicas de las moléculas, como es el caso del coeficiente de difusión asociado a cada subclase de lipoproteína. A partir de la medición de los coeficientes de difusión se calculan directamente los tamaños de las diferentes subclases de lipoproteínas a través de la ecuación de Stokes-Einstein **[27]**

El análisis del plasma mediante DOSY-RMN en 2D genera un espectro de resonancia complejo del que se puede obtener un grado de información superior al que se obtiene en los análisis tradicionales.

Cabe destacar que la medida directa del tamaño de las lipoproteínas es de particular importancia ya que se utiliza para calcular el número de partículas de lipoproteínas dividiendo el volumen espacial de las moléculas de lípidos totales por el volumen medio (es decir, tamaño) de las partículas de lipoproteínas. Por tanto, la RMN en 2D que permite calcular directamente los tamaños de las lipoproteínas produce determinaciones más precisas de las concentraciones de partículas lipoproteicas que los métodos basados en RMN en 1D.

Actualmente, el test Liposcale® ha sido adaptado a las necesidades de la comercialización en cuanto a tiempo y costo mediante el escalado del test en dos dimensiones iniciales, a un solo gradiente optimizado de tal manera que los parámetros determinados reproduzcan el experimento inicial, fijando algunos parámetros del estudio de un conjunto de más de 600 muestras, como por ejemplo el tamaño particular de cada subclase de lipoproteínas y la posición de las funciones analÍticas con las que se deconvoluciona el espectro. Por lo tanto, es una prueba que analiza espectros 1D basándose en la prueba inicial que utilizaba espectros 2D.

2. Contexto del proyecto y objetivos

Biosfer Teslab, una empresa situada en Reus (Tarragona), surgió como spin-off de la Universitat Rovira i Virgili (URV) y el Institut de Recerca Sanitària Pere Virgili (IISPV) con la finalidad de reducir el tiempo necesario entre la obtención de los resultados científicos y su futura aplicación a los pacientes a través del asesoramiento de personal cualificado correspondiente **[28]**. Actualmente, la compañía ofrece servicios analíticos in vitro para monitorizar y estudiar las alteraciones del metabolismo a partir de muestras de biofluidos, tanto humanas como animales, dando paso así a un diagnóstico lo más certero posible y útil para que los profesionales de la salud y a la comunidad científica para avanzar en el conocimiento de la salud.

El test que da origen a la compañía es el test Liposcale®, que es un test avanzado basado en resonancia magnética nuclear (RMN) que permite obtener el perfilado avanzado de lipoproteínas, que determina de manera directa y rápida el tamaño, la composición lipídica y el número de partículas de las principales clases de lipoproteínas (VLDL, LDL y HDL), así como la concentración de partículas de nueve subclases diferentes.

El test Liposcale® se realiza a partir de datos que son el resultado de un análisis de RMN de muestras de suero o plasma sanguíneo humano. Este trabajo surge a razón de la necesidad existente de comprender y prevenir las enfermedades cardiovasculares mediante el análisis del perfil lipoprotéico a partir de muestras de suero o plasma sanguíneo animal. El primer paso para la realización del perfilado avanzado de lipoproteínas en modelos animales es conocer la concentración de colesterol y triglicéridos totales en muestras de suero animal, para posteriormente obtener el resto de parámetros del perfil avanzado de lipoproteínas (lo que sería equivalente al test Liposcale® que la compañía ya comercializa para muestras humanas). Hasta la fecha, la compañía realiza la caracterización avanzada de lipoproteínas en modelos animales haciendo uso de la determinación bioquímica mediante métodos tradicionales (kits enzimáticos), gracias a los cuales se obtiene el colesterol y los triglicéridos totales de las muestras. Dado que los modelos animales presentan unos rangos diferentes de concentración tanto de colesterol como de triglicéridos, no se pueden aplicar los métodos que la compañía utiliza actualmente en el test de muestras humanas. Por lo tanto, el **objetivo** de este trabajo es la **cuantificación de colesterol y triglicéridos totales a partir de espectros de resonancia magnética nuclear utilizando modelos de predicción**. De esta manera, la compañía podrá realizar el perfilado avanzado de lipoproteínas sin la necesidad de realizar una determinación por métodos bioquímicos y, por lo tanto, se podrá realizar el test avanzado de lipoproteínas únicamente mediante el uso de la RMN, tal y como se hace en el test Liposcale® para muestras humanas. Una vez obtenidos estos modelos de predicción se podrán implementar en la compañía para la caracterización avanzada de lipoproteínas en muestras de modelos animales (esta segunda fase no se incluye en el presente trabajo). Las muestras que se usaron en el proyecto pertenecían a tres clientes diferentes y fueron analizadas en el laboratorio previamente a la realización de este trabajo.

Para lograr cuantificar el colesterol y los triglicéridos totales, se plantean los siguientes objetivos específicos:

1. Aplicar y analizar, mediante el uso de Matlab, varias técnicas estadísticas como estadística descriptiva univariante, PCAs, STOCSYs y modelos de regresión de mínimos cuadrados parciales o PLS que permitirán la obtención de un modelo de predicción.
2. Obtener modelos de predicción con fiabilidad suficiente (R = 0.8).
3. Observar con qué preprocesado se obtienen mejores resultados.

Este trabajo ha resultado, personalmente, de gran interés y ha sido clave en la toma de decisiones en relación a mi futuro profesional, ha sido una experiencia que me ha abierto un rango de oportunidades que desconocía resultando en uno de los campos más cautivadores para mí.

3. Metodología

3.1. Muestras analizadas

Para realizar este proyecto se han usado 3 conjuntos de muestras de cerdo sumando un total de 230. Estas muestras llegaron a la empresa Biosfer Teslab para estudiar el perfilado avanzado de lipoproteínas basado en Resonancia Magnética Nuclear (RMN), ya que, como se ha mencionado en la introducción los cerdos son excelentes modelos para estudiar los comportamientos fisiológicos en humanos. Por motivos de protección de datos no se pueden mencionar los nombres reales de las muestras, así que para referirnos a ellas se les llamará muestras o conjuntos 1, 2 y 3 (**Tabla 3-1)**.

Tabla 3-1. Número de muestras con las que se va a trabajar.

Conjunto	Número de muestras
Conjunto 1	19
Conjunto 2	116
Conjunto 3	95

Además de estas muestras, también se dispone de los valores de colesterol total y triglicéridos totales obtenidos mediante métodos bioquímicos estándares.

3.2. Preparación y análisis de las muestras

La primera fase del proyecto es el análisis de las muestras para que se obtengan los resultados con los que se va a trabajar en este proyecto.

El protocolo usado en los laboratorios de la empresa Biosfer Teslab cuenta con reactivos para las muestras, en este caso se usa agua deuterada de Euroisotop, sales NaH_2PO_4YNa_2PO_4 de Labmek, que se usaron para preparar un tampón fosfato salino o PBS (por sus siglas en inglés Phosphate Buffered Saline), como últimos reactivos se usaron NaOH de POCH que permitirán ajustar el pH del tampón.

Para preparar las muestras se descongelaron y se homogeneizaron mediante un robot manipulador de líquidos Gilson. Posteriormente se alicuotaron 50 µL de muestra en el tubo de NMR y se

mezclaron con 150 µL de un pool de suero bajo en lípidos (colesterol total < 200 mg/dL y triglicéridos totales < 100 mg/dL). A continuación, se diluyeron las muestras con 300 µL de PBS 50 mM (pH = 7,4) y 50 µL de agua deuterada (D2O), finalmente las muestras se homogenizan en el tubo de NMR.

Para concluir con el análisis de las muestras diluidas se aplica la espectrometría de resonancia magnética nuclear de protones (1H-RMN) utilizando un espectrómetro Bruker Avance III a una frecuencia de 500 MHz. Los espectros se obtuvieron mediante dos experimentos de NMR. A cada muestra se les aplicó dos pulsos: el pulso NOESY (por sus siglas en inglés Nuclear Overhauser Effect SpectroscopY) cuyo espectro se analiza con finalidad de control de calidad del espectrómetro y el pulso LED (por sus siglas en inglés Longitudinal Eddy-Current Delay) el cual se utilizó para la realización de este proyecto. Estos pulsos son equivalentes a la resolución que se tiene de los espectros de las muestras siendo el pulso NOESY el que presenta una resolución peor ya que incluye todas las moléculas que se pueden encontrar en una muestra de plasma mientras que el pulso LED ofrece una mejor resolución porque solo permite visualizar las macromoléculas como el colesterol y los triglicéridos contenidos dentro de las lipoproteínas.

3.3. Métodos estadísticos

A lo largo de todo el proyecto han sido necesarios varios métodos estadísticos que se detallan en esta sección. Primeramente se hizo un análisis exploratorio de los datos, para observar la distribución de las variables y el comportamiento de cada uno de los conjuntos del estudio. Para esta primera parte se realizó estadística descriptiva y un análisis de componentes principales con la finalidad de tener un mayor conocimiento de los conjuntos que queríamos estudiar. En la segunda parte donde se realizaron STOCSY para seleccionar las regiones del espectro que iban a ser utilizadas en los modelos de predicción tanto para la predicción de colesterol total como para la predicción de triglicéridos totales. Más adelante se verán ejemplos gráficos en relación con estas técnicas.

3.3.1. Estadística descriptiva

Primero se usó la estadística descriptiva univariante para conocer mejor las muestras, se realizó un estudio de la media, la mediana, los máximos y mínimos seguido del cálculo de los rangos, los cuartiles y la desviación estándar, toda esta información proporcionará una idea de que se puede esperar de dichas muestras y cuáles son comparables con cuáles. Con la finalidad de observar la distribución de las muestras se realizaron histogramas para todos los conjuntos tanto para colesterol como para triglicéridos.

3.3.2. Análisis de Componentes Principales

La siguiente técnica estadística que se usó fue el PCA o Análisis de Componentes Principales **[29]**. Esta técnica estadística multivariante permite simplificar la complejidad de espacios muestrales a la par que permite visualizar el comportamiento de los conjuntos, también agrupa la mayor variación posible con la finalidad de reducir el número de variables. Para realizar un PCA se deben introducir los datos del colesterol y los triglicéridos totales que nos interesa observar además de las 23 variables que predice el test Liposcale®, estas son: las concentraciones de colesterol y triglicéridos de VLDL, LDL, IDL y HDL, el número total de partículas de colesterol y triglicéridos para VLDL,LDL y HDL, las concentraciones de dichas partículas, menos las IDL, según su tamaño (grande, mediano y pequeño) y el tamaño medio de las partículas grandes, pequeñas y medianas. Los resultados que se obtienen se representan en una gráfica de Scores del PCA donde se puede apreciar visualmente el modo en que se agrupan las muestras. Aquellas que se encuentran más juntas en el gráfico de scores significa que presentan comportamientos metabólicos similares.

Una vez obtenidos los *Scores* se debe mirar lo que se conoce como gráfico de *Loadings*. Este gráfico explica la relevancia que tienen las variables en los componentes principales, como en este caso estamos representando los datos en dos ejes, representamos los *loadings* para estos dos componentes principales. Brevemente, los *loadings* indican cuáles son las variables que tienen más relevancia, tanto valores positivos como negativos y que provocan una separación de las muestras en el gráfico de *scores* acorde al signo de la influencia que ejercen.

3.3.3. STOCSY

El STOCSY es una representación gráfica de las correlaciones entre los espectros de RMN y la variable de predicción que queremos estudiar (en nuestro caso colesterol y triglicéridos) **[30]**. Esta técnica representa el coeficiente de correlación que se obtiene para cada una de las regiones del espectro con la variable que queremos predecir. Es por este motivo que se usa un rango de 1, siendo la máxima correlación positiva (coeficiente de correlación R = 1) hasta -1 que equivale a la máxima correlación negativa (R = -1). El eje x representa el desplazamiento químico como ppm, estos son una conversión de unidades con el fin de mantener la escala sea cual sea el espectrómetro que se usa para analizar las muestras, se mide el desplazamiento químico y se divide entre la frecuencia del espectrómetro, este proceso es necesario para evitar valores de ordenes demasiado elevados.

Este método se usó en este proyecto para seleccionar las regiones que presentaban una mayor correlación con las variables que se quiere predecir (colesterol y triglicéridos) y que por lo tanto serán las variables de entrada para el modelo de predicción. De este modo reducimos la cantidad de variables de entrada en el modelo para que dichas variables expliquen las variables a predecir y que por lo tanto presenten una mayor correlación. En este caso concreto las variables generadas corresponden con las diferentes partes de las macromoléculas (lipoproteínas), como los grupos metilo, para así poder seleccionar qué regiones presentan una correlación mayor (haya mayor número de esas moléculas).

3.3.4. Regresión de Mínimos Cuadrados Parciales

Para la realización de los modelos PLS, en el presente trabajo se han evaluado dos tipos de preprocesado: el Mean Centering o centrado de media y el Autoscale o autoescalado, además se han realizado modelos PLS para el colesterol y para los triglicéridos por separado.

La función del primero es sustraer la media de todas las observaciones para una variable determinada y así obtener una nueva media que sea cero; de esta forma ningún conjunto tiene más peso que otro **[31]**. Por otro lado, para hacer un autoescalado hay que realizar un centrado de media para cada variable y luego dividirla entre la desviación estándar de la variable correspondiente, proporcionando así igualdad entre las variables **[32]**.

El método que se explica a continuación (modelo PLS o modelos de regresión de mínimos cuadrados parciales) es el método principal que se ha utilizado este proyecto ya que va a permitir cuantificar la concentración de colesterol y triglicéridos totales en las muestras de modelos animales, a partir de

los cuales permitirá en un futuro proyecto obtener la caracterización avanzada de lipoproteínas en dichos modelos.

Los modelos PLS son unos métodos estadísticos que dan una estimación para tratar datos colineales **[33]**. Se usa comúnmente al trabajar con espectroscopía donde uno de los objetivos es predecir la composición química de un espectro. En el presente trabajo, al crear modelos PLS, no se contempla la interacción entre variables.

Primeramente se entrena al modelo (conjunto de entrenamiento o training set) dándole las regiones de los espectros seleccionadas en el STOCSY y posteriormente se validan (conjunto de validación o validation set) los resultados con muestras totalmente nuevas que no ha visto anteriormente. Para la realización de esta validación se aplicó un método de validación conocido como Venetian Blinds, el cual ofrece el número óptimo de variables latentes para usar en el modelo PLS que se está estudiando **[34]**. Una variable latente es una variable inferida que se ha obtenido mediante modelos matemáticos usando otras variables. En el presente proyecto las variables latentes se corresponden con una constante que selecciona el mismo programa para aplicar una influencia a las variables de entrada, estas son las regiones de interés previamente mencionadas.

3.4. Programario

A continuación, se exponen los programas informáticos que se han usado para realizar este proyecto que engloba desde la creación de una base de datos hasta la creación de los modelos PLS.

3.4.1. Excel

Para poder crear y trabajar con los datos se creó una base de datos en Microsoft Office Excel 2007. Este programa también se ha usado en este trabajo con la finalidad de poder representar en una tabla los datos estadísticos descriptivos y comparación de resultados.

3.4.2. Minitab

Minitab 18 (2018) se ha usado para la creación de gráficos de cajas y para corroborar los resultados obtenidos en MATLAB y Excel® mediante una comparación directa de los resultados obtenidos, aunque como programa ofrece un amplio abanico de posibilidades a la hora de hacer cálculos estadísticos o visualizar gráficamente los resultados que se obtienen.

3.4.3. MATLAB

Finalmente, el centro de todo el proyecto ha consistido en el uso de MATLAB versión 7.10.0.499 R2010a (MathWorks). Este programa es el que ha calculado la estadística descriptiva a partir de los datos de las muestras recopiladas en la base de datos, los cuales se importaron a dicho programa una vez se acabaron de tratar, mediante la programación se ha creado un código y usado las funciones correspondientes para la obtención del PCA, los STOCSY y por último los PLS.

Para crear los modelos PLS, en este trabajo, se ha entrenado al programa dándole el 70% de las muestras con las que se trabaja. El 30% restante de muestras se usa para hacer la prueba de predicción y cuantificación (validación), para evaluar la robustez del modelo se obtiene el valor de la *R* indicando el coeficiente de correlación entre los datos obtenidos y los datos reales , cuanto más afín sean los valores predichos a la línea de predicción mayor será la *R*, siendo *R* = 1 al obtener el valor máximo en una predicción perfecta.

Se puede intentar reducir el error al seleccionar manualmente las regiones de interés para añadirlas al modelo usando una función, que se añade a MATLAB®, llamada “getrang”, esta permite al usuario seleccionar una región del espectro de las muestras que más le interesen y luego aplicando la función de “include” podemos añadir esas regiones al modelo PLS y, de esta forma, eliminar del modelo todo aquello que tenga correlaciones nulas o sin interés para el proyecto. Para saber si una región es de interés o no se puede usar los STOCSY que nos indican qué tipo de correlación hay en ellas. Para este trabajo se han usado las regiones relacionadas con los grupos metilo por su alta relación con las lipoproteínas, estas regiones se encuentran en 0,8 ppm, 1,2 ppm, 2,1 ppm, 3,6 ppm y 5,7 ppm de lo que se conoce como H *chemical shift* (eje *x*) o desplazamiento químico, no intensidad como se suele ver en otros casos, que permite una estandarización para equiparar los resultados de cualquier espectrómetro que se use.

4. Resultados y discusión

Para contextualizar al lector, a continuación se habla de los niveles de colesterol y triglicéridos que son considerados normales tanto en humanos como en cerdos. No hay mucha información sobre el rango de colesterol y triglicéridos que es habitual en cerdos en internet, pero sí se indica que son siempre un poco superiores en niveles de colesterol al de los humanos cuando se les alimenta de forma ‘’sana’’. Un estudio publicado en la Revista Citecsa de Colombia muestra el efecto de la palma africana en la dieta para cerdos **[35]**, en este estudio los cerdos se dividen en tres grupos: uno que es alimentado con alimento balanceado comercial 100%, el segundo grupo que se alimenta con una dieta balanceada comercial del 90% y un 10% de fruto de palma, al último grupo se le da 100% alimento balanceado comercial más fruto entero de palma a voluntad del cuidador.

El grupo control en este caso es el primero y según sus datos el valor de colesterol en cerdos rondaría los 137,5 mg/dL y el de triglicéridos alrededor de 50 mg/dL pero como se ha mencionado antes no hay mucha información con la que contrastar estos datos.

Para hacerlo más comprensible, en humanos sanos, el rango de colesterol oscila entre 125 mg/dL a 200 mg/dL tanto para hombres como mujeres mayores de 20 años **[36, 37]**. Para edades inferiores los valores son de concentraciones menores a los mencionados. Respecto a los niveles de triglicéridos deben ser menores de 150 mg/dL para que sean considerados normales **[38, 39]**.

Como se ha mencionado antes esta información está presente para dar a conocer cuales serían los niveles esperados de colesterol y triglicéridos en cerdos y en humanos y así aportar un punto de referencia a los lectores.

4.1. Estadísticos descriptivos

Como se menciona anteriormente, el primer paso a realizar en el estudio después de la creación de la base de datos es la estadística descriptiva, cuyo objetivo es conocer los rangos de las variables que se quieren estudiar con la finalidad de poder interpretar los resultados que se obtendrán posteriormente en los modelos de predicción. En la **Tabla 4-1** y **4-2** se muestran los resultados obtenidos de la estadística descriptiva tanto para el colesterol total como para triglicéridos totales respectivamente a partir de los datos de bioquímica, de la base de datos, juntamente con los diagramas de cajas correspondientes que se muestran en las **Figuras 4-1** y **4-2**.

Tabla 4-1. Resumen de los estadísticos descriptivos para el colesterol total en mg/dL.

Nombre	Conjunto 1	Conjunto 2	Conjunto 3
Nº de muestras	19	116	95
Media CT (mg/dL)	336,49	250,87	78,02
Mediana CT (mg/dL)	302,78	106,15	76,95
Mínimo CT (mg/dL)	85,07	54,91	57,62
Máximo CT (mg/dL)	668,21	715,39	111,76
Q1 CT (mg/dL)	128,58	80,43	70,86
Coef. Var CT (mg/dL)	61,65	79,50	14,33
Q3 CT (mg/dL)	544,18	423,63	85,27
Desv. Est. CT (mg/dL)	207,45	199,45	11,18

Podemos observar que los tres conjuntos presentan medias bastante diferentes, pudiendo decir que el conjunto 1 y conjunto 2 son hipercolesterolémicos y el conjunto 3 presenta unos valores de colesterol más bajos. De la misma manera podemos ver que existe una mayor dispersión en los dos primeros conjuntos, debido a que el rango de valores para el colesterol del conjunto 3 es mucho más estrecho, pudiéndose considerar a las muestras del tercer conjunto ligeramente hipocolesterolémicas.

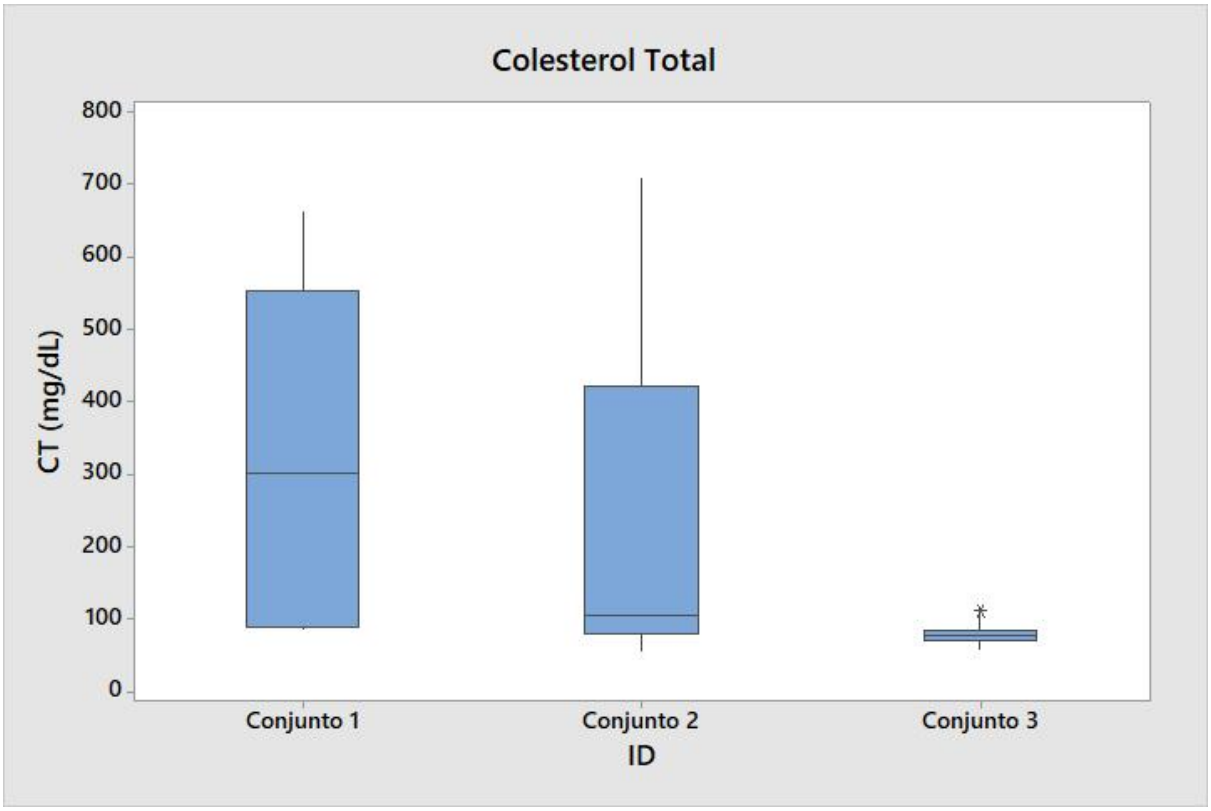


Figura 4-1. Diagrama de cajas correspondiente al colesterol total.

La figura anterior (**Figura 4-1**) es el diagrama de cajas de los 3 conjuntos, y a su vez se puede interpretar como un resumen visual de los resultados de la estadística descriptiva donde se aprecian los rangos del colesterol total. Observando los datos proporcionados, el conjunto 1 en particular es el que tiene los niveles de colesterol más elevados de media, en contraste, el conjunto 3 difiere drásticamente en el tamaño del rango, es decir, los valores de los datos son mucho más pequeños que en los otros dos conjuntos. Tal y como se ha mencionado antes, el comportamiento del conjunto 1 refleja que las muestras tienen mayoritariamente características hipercolesterolémicas mientras que el conjunto 3 muestra un comportamiento ligeramente hipocolesterolémico. En relación con el conjunto 2, las muestras presentan niveles mayoritariamente normales de colesterol (ya que presenta su mediana en un valor ligeramente superior al 100 mg/dL) aunque en algunas muestras se refleja una tendencia hipercolesterolémica. Hay que considerar el número de muestras de cada conjunto ya que la diferencia es significativamente elevada entre la cantidad de muestras del conjunto 1 y el resto de los conjuntos.

Tabla 4-2. Resumen de los estadísticos descriptivos para triglicéridos totales en mg/dL.

Nombre	Conjunto 1	Conjunto 2	Conjunto 3
Nº de muestras	19	116	95
Media TG (mg/dL)	61,02	44,12	30,92
Mediana TG (mg/dL)	54,92	38,97	29,23
Mínimo TG (mg/dL)	12,40	5,31	8,86
Máximo TG (mg/dL)	113,37	133,74	107,17
Q1 TG (mg/dL)	35,65	26,13	19,71
Coef. Var TG (mg/dL)	51,32	58,39	45,72
Q3 TG (mg/dL)	83,48	61,12	38,75
Desv. Est. TG (mg/dL)	31,32	25,76	14,13

Los resultados que se han obtenido de la estadística descriptiva referentes a los triglicéridos totales cuentan con valores cercanos a lo que se considera normal, se puede ver cómo las muestras siguen el mismo comportamiento que para el colesterol.

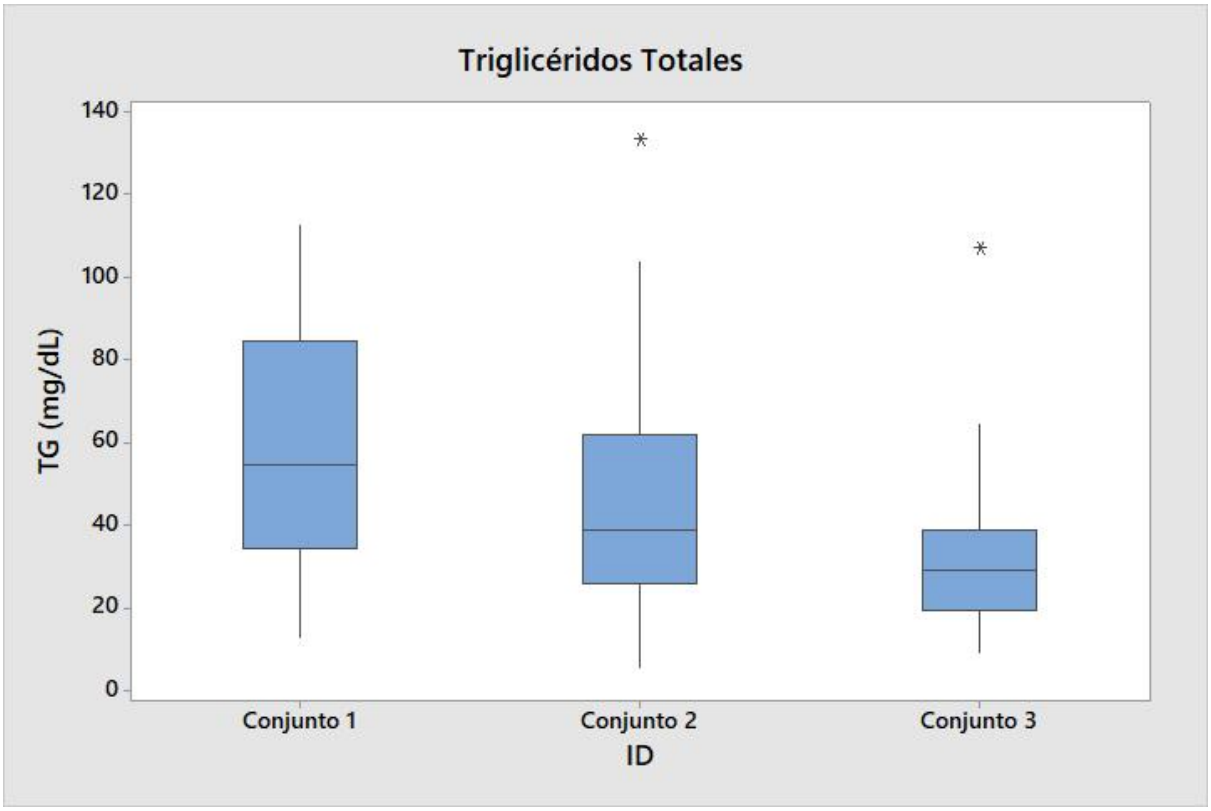


Figura 4-2. Diagrama de cajas correspondiente a los triglicéridos totales.

En la figura anterior (**Figura 4-2**) se muestran los diagramas de cajas correspondientes a los tres conjuntos para los valores de triglicéridos. Tal y como se expresa anteriormente, los niveles de triglicéridos en cerdos sanos rondan los 50 mg/dL por lo que el conjunto 2 tendría los resultados más cercanos a los niveles de normalidad en triglicéridos. Contrariamente, encontramos que el conjunto 1 podría corresponder con valores ligeramente hipertrigliceridémicos y que las muestras del conjunto 3 corresponderían con niveles ligeramente hipotriglicémicos. A diferencia de lo que sucede con el colesterol, los datos obtenidos para los triglicéridos son mucho más parecidos. Nuevamente se debe recordar que el número de muestras entre el conjunto 1 y los otros dos es significativamente diferente.

Aquellos datos atípicos que aparecen en ambos diagramas de cajas se han mantenido en el estudio debido a que las muestras deben ser tratadas como si pertenecieran a casos humanos, por ende, siempre existe la posibilidad de que aparezcan casos extremos en humanos que deban ser estudiados. Se debe tener en cuenta que al tratarse de seres vivos, los datos son muy variados y que se necesita la máxima representación que sea posible de todos los casos para que el modelo presente la mayor fiabilidad en el diagnóstico clínico.

Para poder ver la distribución de las variables a estudiar se realizaron histogramas, que aparecen representados en el **Anexo I**. En ellos se puede observar como ninguno de los conjuntos presenta una distribución de los datos estrictamente normal. Esto, muy probablemente, se deba a que las poblaciones no han sido elegidas aleatoriamente sino que han sido seleccionadas expresamente por seguir una dieta que les ha inducido cambios en el perfilado lipídico interesantes para estudiarlos.

4.2. PCA

El siguiente paso consiste en realizar un PCA para los conjuntos. Esto nos mostrará su distribución según las variables de Liposcale®, posteriormente se representarán en un gráfico de *scores,* donde se seleccionarán dos componentes principales que explican el 63,71% de la varianza de las muestras. En la **Figura 4-3** se muestra el PCA obtenido con los datos del proyecto.

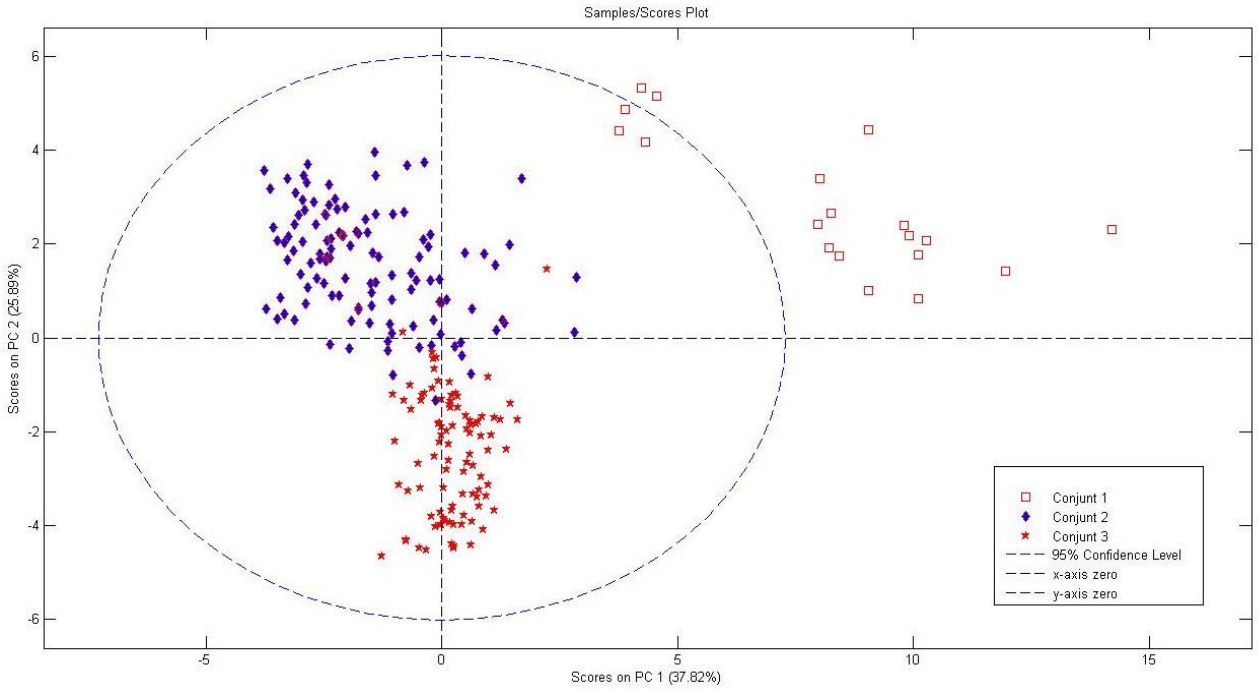


Figura 4-3. Scores PCA de los tres conjuntos. Arriba niveles altos de colesterol y abajo valores menores de colesterol. La izquierda indica una concentración menor de triglicéridos y la derecha indica una concentración mayor de triglicéridos.

A partir del gráfico de scores se puede ver que el componente principal 1 (PC1 en el eje *x*) está separando principalmente el conjunto 1 de los otros dos. En cambio, el PC2 está separando el conjunto 1 y 2 del conjunto 3. Para saber qué variables están influenciando es esta separación se realizan los gráficos de *loadings*, que permitirán de las 23 variables que devuelve el test Liposcale®, saber cuáles son las más influyentes para cada componente principal.

El PC1 o componente principal 1 está altamente influenciado por las partículas LDL-C(colesterol), las LDL-TG (triglicéridos) y sus concentraciones sobre todo para las partículas que son grandes y por último las partículas HDL-C, en particular aquellas que disponen de un tamaño mayor. Por lo tanto, las muestras del conjunto 1 presentan valores incrementados de estas variables, lo que es indicativo de que la separación que se está realizando en los scores corresponde con los niveles de LDL y, por ende, el eje *x* está relacionado con el colesterol.

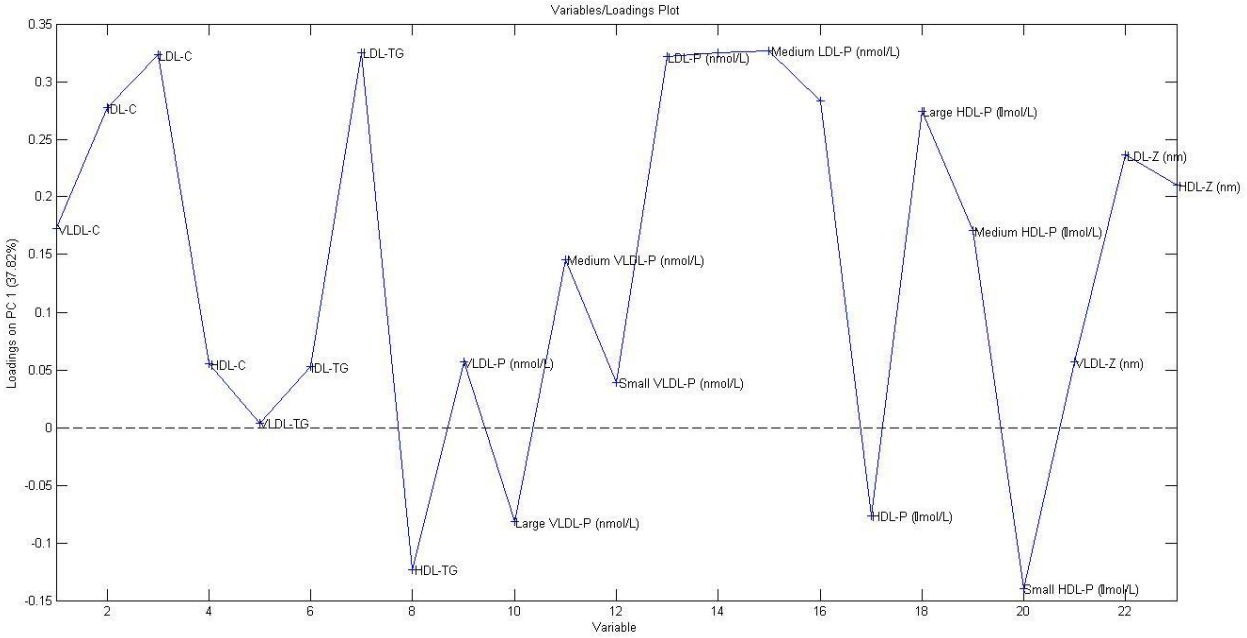


Figura 4-4. Loadings del componente principal 1 en el PCA.

Por otro lado, en los *Loadings* del segundo componente principal o PC2 las partículas que tienen más influencia son las VLDL-C (colesterol) y VLDL-TG (triglicéridos) para todos los tamaños de estas partículas. En menor medida, las partículas HDL-TG también son influyentes por lo que se puede intuir que los conjuntos 1 y 2 presentan niveles de estas variables superiores a los del conjunto 3. Esto indica que la separación que se lleva a cabo en el eje *y* corresponde con los triglicéridos ya que las partículas VLDL son las que transportan principalmente los triglicéridos por el torrente sanguíneo.

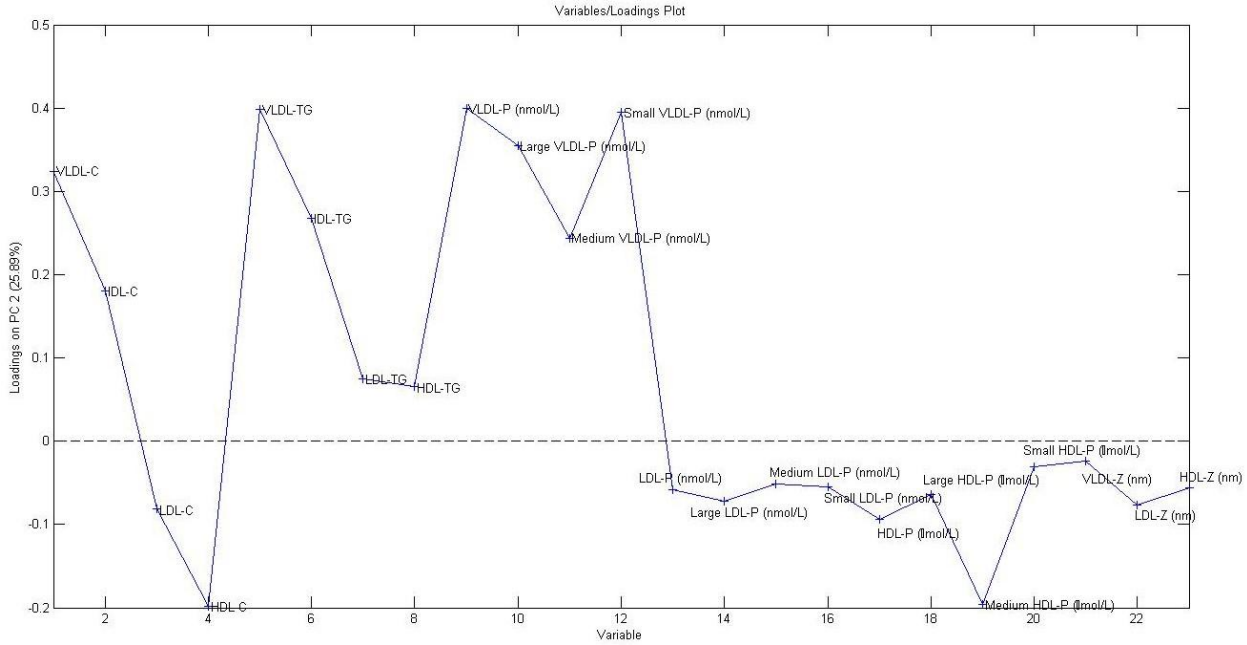


Figura 4-5. Loadings del componente principal 2 en el PCA.

Tras haber realizado la primera parte del trabajo se decidió excluir a las muestras del conjunto 3 debido a sus grandes diferencias en rangos con las muestras de los otros conjuntos. Este conjunto presenta unos rangos de valores poco comparables con el resto de muestras por lo que puede afectar negativamente al objetivo del trabajo. Esta decisión está basada en la observación de los resultados de la estadística descriptiva donde se expone claramente que los valores para el colesterol total de estas muestras son predominantemente inferiores al resto.

4.3. STOCSY

Como se ha mencionado previamente este método estadístico se usa para observar las correlaciones entre las regiones de los espectros de RMN de las muestras y las variables de predicción. Se realizaron STOCSY para colesterol y triglicéridos para cada conjunto y también uno para todas las muestras juntas tanto de triglicéridos como para colesterol.

Siguiendo el orden anterior encontramos primero los resultados gráficos del colesterol (**Figuras 4-6, 4-7** y **4-8)**. Como se puede observar en la **Figura 4-6** que corresponde con las muestras del conjunto 1, el colesterol total tiene muy buena correlación en las zonas de interés que son aquellas regiones relacionadas con los grupos metilos mencionadas anteriormente en la metodología (0,8 ppm, 1,2 ppm, 2,1 ppm, 3,6 ppm y 5,7 ppm del eje *x*). La zona con la mayor intensidad que aparece con correlaciones nulas es la supresión del agua que sale alrededor de 4,7 ppm y su aparición no es influyente para el experimento.

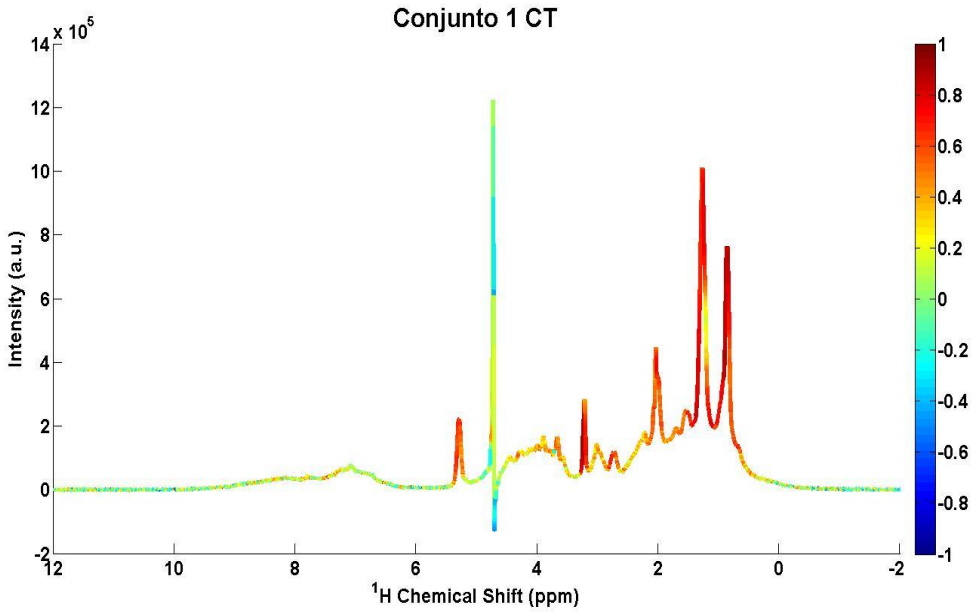


Figura 4-6. STOCSY correspondiente al conjunto 1 para el colesterol total.

Para el conjunto 2, mostrado en la **Figura 4-7**, se observa un comportamiento que presenta menor correlación en comparación al observado previamente, no obstante, las regiones de interés siguen presentando las correlaciones más altas del espectro.

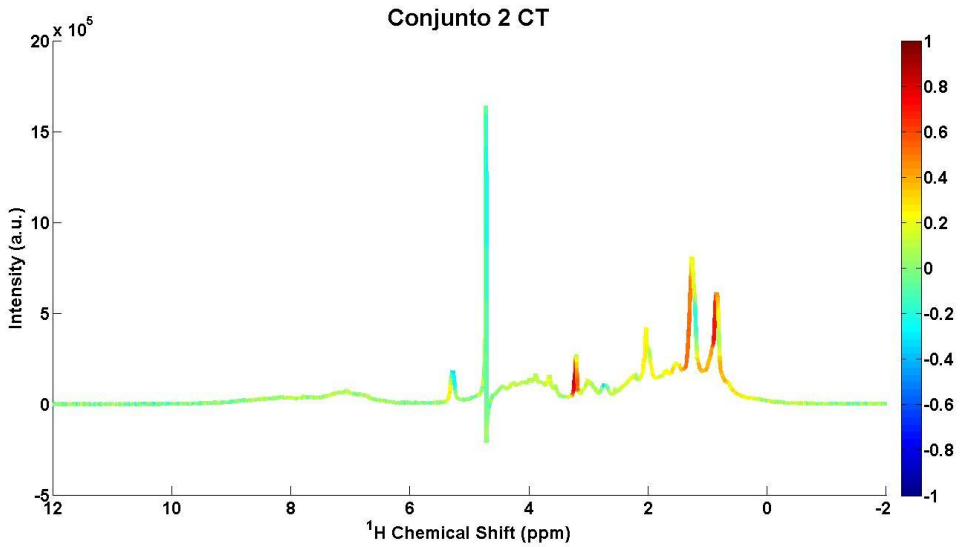


Figura 4-7. STOCSY correspondiente al conjunto 2 para el colesterol total.

Para finalizar con el colesterol, tal y como se ha mencionado antes se realiza un STOCSY de los dos conjuntos, el resultado mostrado en la **Figura 4-8** indica una correlación diferente a la de los dos conjuntos por separado. Este STOCSY se comporta de manera similar al del conjunto 2 ya que este posee la mayoría de los datos.

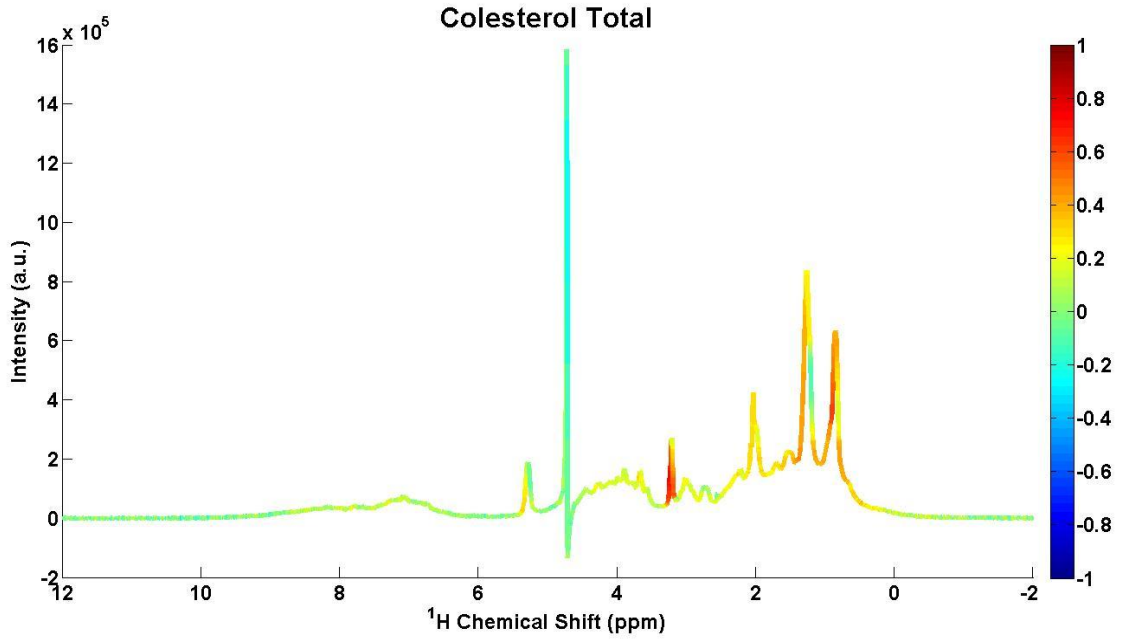


Figura 4-8. STOCSY correspondiente a los tres conjuntos juntos para el colesterol total.

Una vez mostrados y explicados los resultados de los STOCSY para el colesterol total y sus principales características hay que hacer lo mismo para los triglicéridos totales. Teniendo en cuenta los resultados de la estadística descriptiva, cabe esperar que los resultados de los STOCSY tengan un mejor comportamiento para triglicéridos que el que han tenido para colesterol ya que los rangos entre conjuntos son bastante comparables. A esto hay que añadir que la varianza es mucho menor en triglicéridos que en colesterol.

Tal y como sucedía con el colesterol total, el conjunto 1 presenta una gran correlación positiva para las regiones de interés del espectro como se puede ver en la **Figura 4-9**, sobre todo en la parte izquierda de los picos.

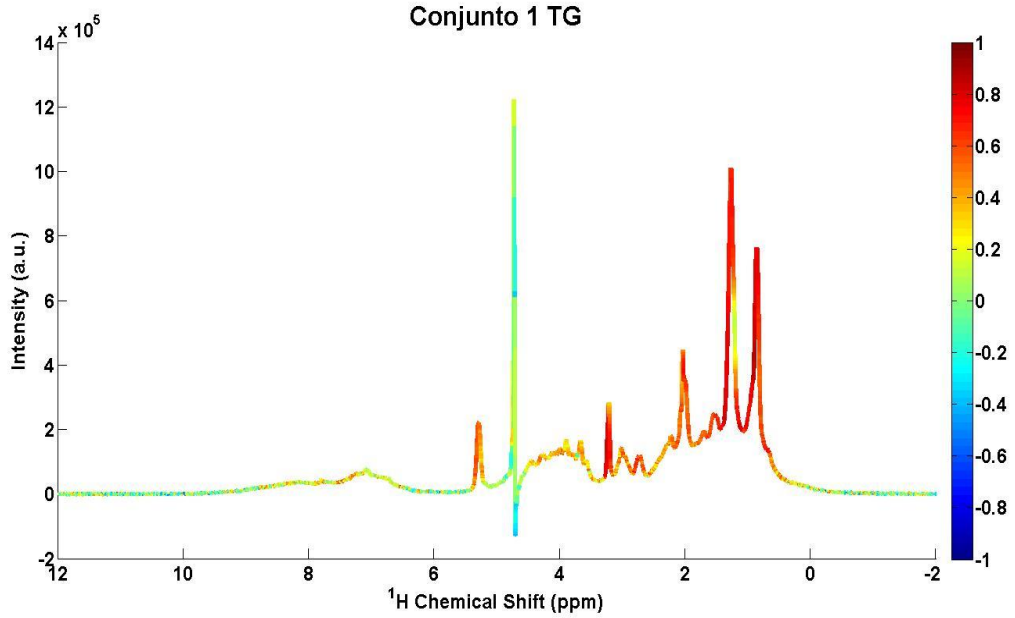


Figura 4-9. STOCSY para los triglicéridos totales del conjunto 1.

Para el conjunto 2 no se encuentran tan buenas correlaciones (**Figura 4-10**) pero resulta ser un modelo interesante ya que la región izquierda del pico presenta una alta correlación que es la zona que se corresponde con la parte donde resuenan más las partículas VLDL y por ende los triglicéridos.

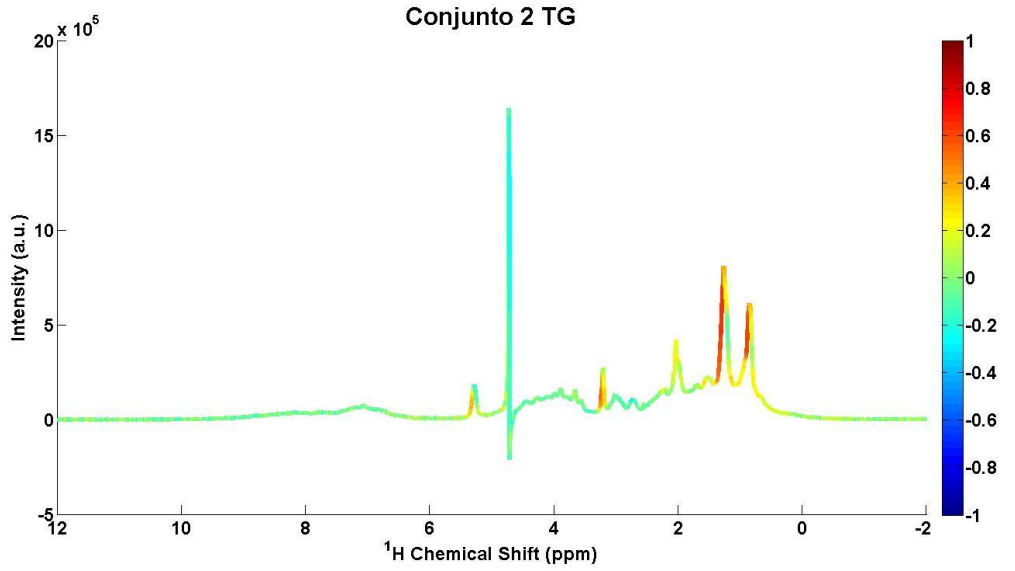


Figura 4-10. STOCSY para los triglicéridos totales del conjunto 2.

Por último, al igual que para el colesterol, se realizó un STOCSY para los dos conjuntos con la finalidad de observar la correlación en triglicéridos. Como se puede ver en la siguiente la correlación (**Figura**

4-11) en las regiones de interés es generalmente positiva, siendo este un modelo relevante para el modelo PLS que se realizará posteriormente.

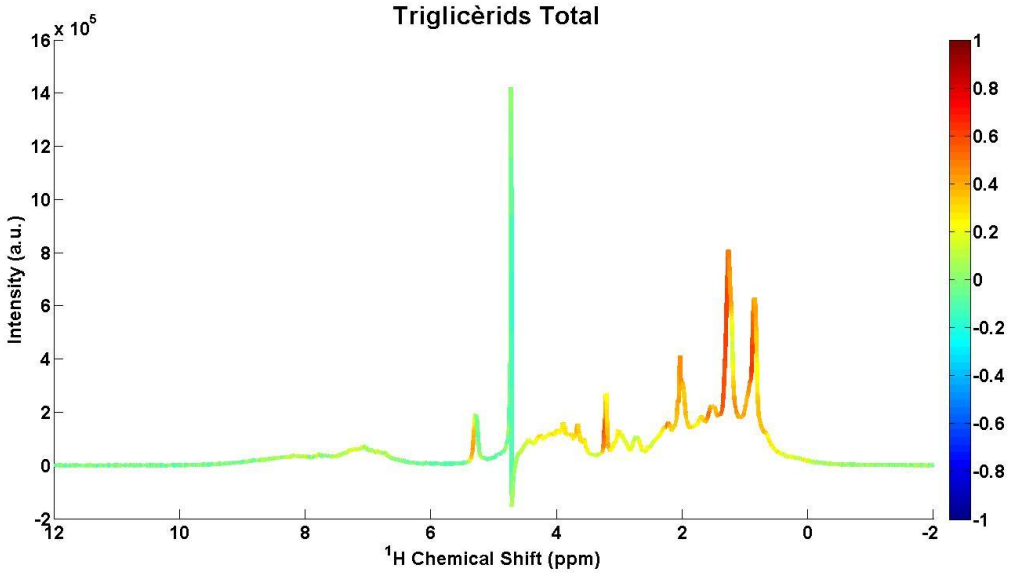


Figura 4-11. STOCSY para los triglicéridos totales correspondiente a los 3 conjuntos juntos.

Como se ha mencionado anteriormente el resultado para los triglicéridos ha resultado ser más favorable que el del colesterol en cuanto a correlaciones pero esto no es definitorio para que vayan a salir mejores modelos PLS para los triglicéridos que para el colesterol.

4.4. Modelos PLS

Una vez finalizada la obtención de STOCSY solo queda un paso para llegar al resultado final de este trabajo, la creación de modelos PLS. Su finalidad, como ya se ha comentado anteriormente, es obtener un modelo que prediga y cuantifique la composición de lipoproteínas de las muestras que se han usado. Para realizar los modelos de predicción se han utilizado dos métodos que se muestran en la siguiente tabla (**Tabla 4-3**).

Tabla 4-3. Resumen de los métodos usados para crear modelos PLS.

Método	Entrenamiento	Prueba
1	70 % de las muestras	30 % de las muestras restantes
2	10 % (valores más altos) + 10% (valores más bajos)+ 50 % (valores aleatorios) de las muestras	30 % de las muestras restantes

Como se observa en la **Tabla 4-3**, el método 1 consistió en separar las muestras en dos grupos de manera aleatoria, el 70% (*N*= 95, donde *N* es el número de muestras) que será dedicado al entrenamiento del modelo para poder obtener unos buenos resultados y el 30% (*N* = 40) restante que servirá como prueba para observar cómo de certero es el modelo o la validación. Estos porcentajes representan muestras seleccionadas al azar para procurar tener una representación de todos los rangos. El método 2 (**Tabla 4-3**) consistió en seleccionar el 10% (*N* = 14) de los valores más altos, el 10% (*N* = 14) de los valores más bajos, el 50% (*N* = 68) al azar y el 30% (*N* = 39) restante para la prueba final al igual que en el otro caso. Toda esta selección se debe hacer tanto para colesterol como para triglicéridos. En el momento de creación de modelos PLS se le indicó a Matlab qué zonas del espectro se iban a usar (las obtenidas a partir del STOCSY) mediante la función 'include'. Esto permitiría al modelo no estar influenciado por zonas donde las variables de interés prácticamente no resonaban y que por lo tanto, no estarían dando información relevante al modelo.

Una vez creados los modelos PLS hay que seleccionar los mejores. Para ello se debe tener en cuenta algunos criterios; el más relevante sería que el valor de la regresión o *R* sea lo más alto posible debido a que este parámetro indica lo robusto y fiable que es el modelo porque nos muestra cuánto se ajustan las predicciones que se obtienen a los datos reales. El siguiente criterio más relevante consiste en observar cuántas variables latentes son necesarias para crear el modelo con un valor de *R* que sea suficientemente elevado. En el presente proyecto las variables latentes se corresponden con

una constante que selecciona el mismo programa para aplicar una influencia a las variables de entrada que son las regiones de interés previamente mencionadas. Para ayudar en la selección de variables latentes se usa la validación de venetian blinds explicada en la metodología. El último criterio para la selección de un buen modelo PLS es qué porcentaje explican las variables latentes sobre el modelo. Podemos considerar que un modelo es mejor si la varianza explicada con pocas variables latentes es alta, esto indica que necesita poca información para poder ofrecer una buena predicción, o bien de colesterol o bien de triglicéridos en este trabajo.

A continuación se muestran los resultados obtenidos para el colesterol y los triglicéridos aplicando el método 1 donde se utilizó 70% y el 30% al azar y después se mostrarán los resultados del método 2 aplicando 10% de valores más altos, el 10% de valores más bajos, el 50% de valores aleatorios y el 30% de valores restantes para la prueba de validación.

Método 1 predicción de colesterol: Las **Figuras 4-12** y **4-13** muestran los resultados de los modelos PLS ordenados según la favorabilidad del modelo siendo el último el más óptimo y el primero el que menos. De los 10 modelos creados se han seleccionado los tres mejores. Tanto la creación de modelos como la selección de estos se realiza a base de prueba y error, es decir, no hay una cantidad definida de modelos a crear o seleccionar. En el presente trabajo se seleccionarán tres de los mejores modelos y no se compararan con los que ofrecen peores resultados por que la empresa no tiene ningún interés real en usar modelos con peor capacidad predictiva.

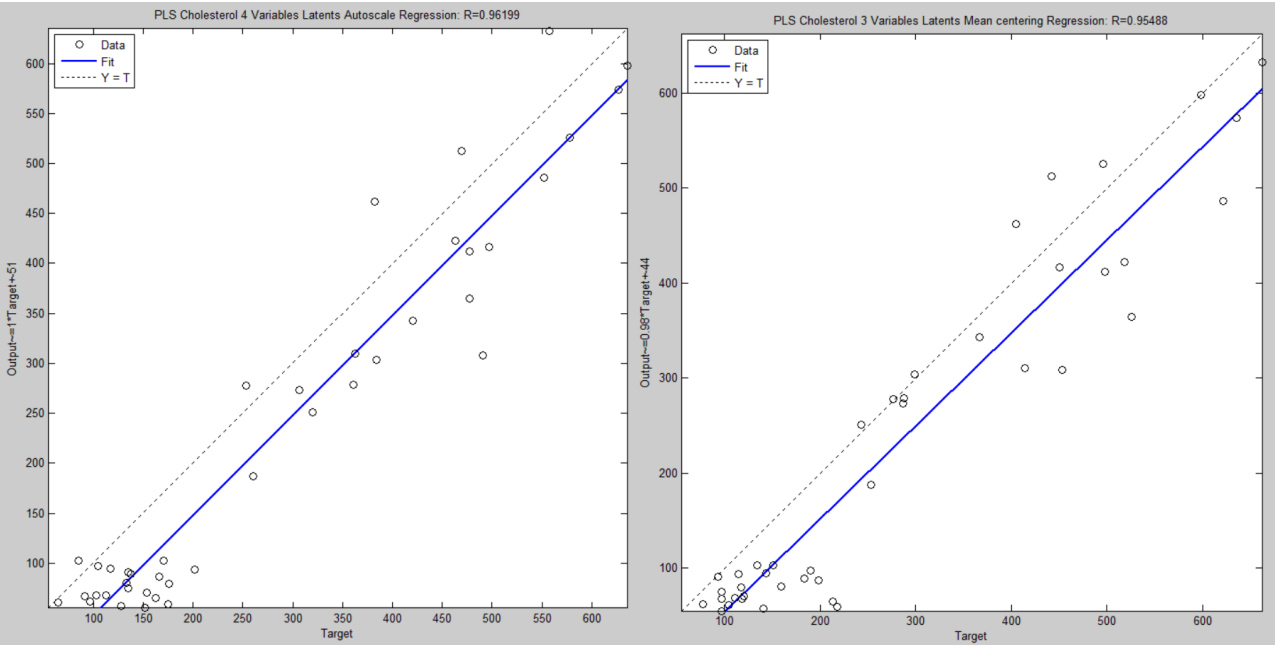


Figura 4-12. Modelo PLS para colesterol con 4 variables latentes, R = 0,96 y preprocesado con Autoscale (izquierda). Modelo PLS para colesterol con 3 variables latentes, R = 0,95 y preprocesado con Mean centering (derecha).

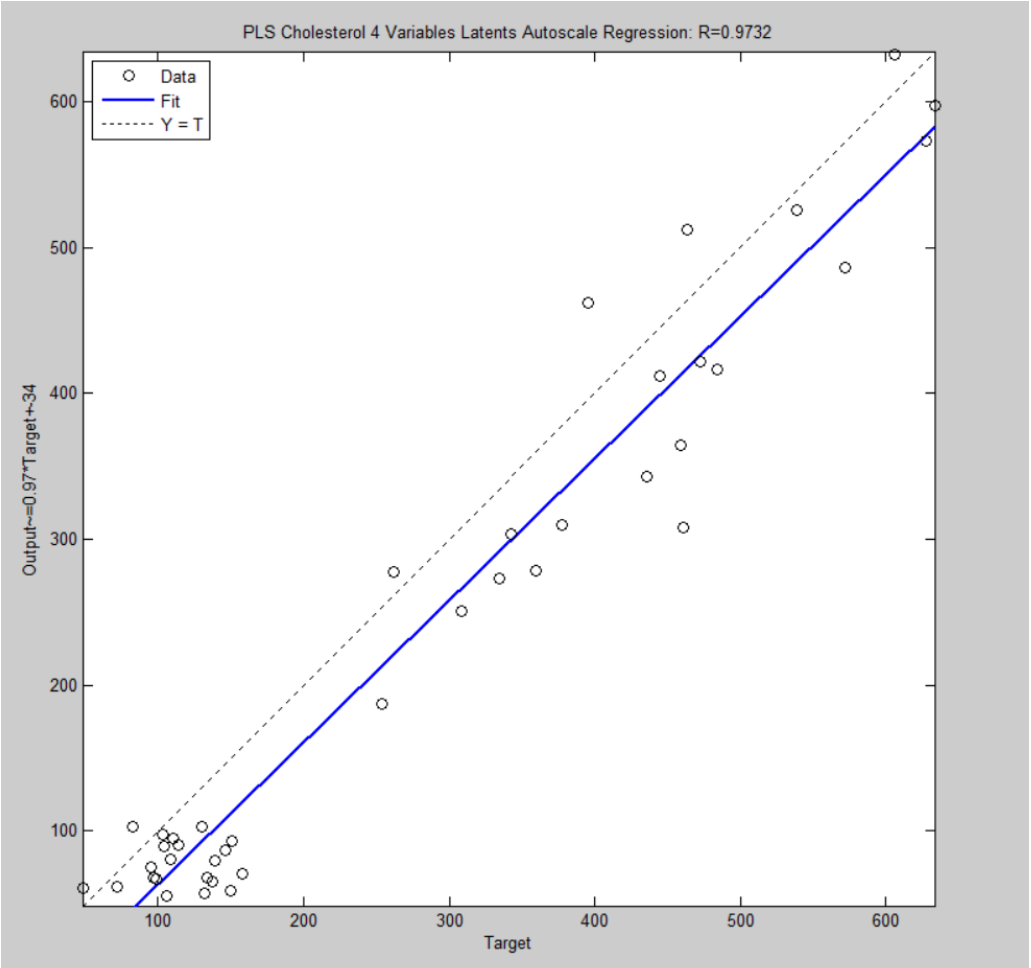


Figura 4-13. Modelo PLS para colesterol con 4 variables latentes, R = 0,97 y con preprocesado Autoscale.

Como se ha mencionado anteriormente hay que seguir unos criterios para seleccionar el mejor modelo, hay que fijarse en la *R*, el número de variables latentes y si es necesario la varianza que explican las variables latentes. Según estos criterios el mejor modelo sería el de la **Figura 4-13** debido a que con 4 variables latentes obtiene una *R* de 0,97 siendo el valor más elevado que se ha obtenido en los modelos creados. En este caso el preprocesado Autoscale garantiza un mejor resultado que el Mean centering.

Método 1 predicción de triglicéridos: Ahora se debe hacer el mismo procedimiento para los triglicéridos. Curiosamente los resultados que aparecen en las **Figuras 4-14** y **4-15** no son mejores que los resultados que se obtuvieron para el colesterol, descartando el hecho de que tener un STOCSY con mayores correlaciones daría una mejor predicción en el modelo PLS. Sin embargo, la razón de estos resultados reside en que el rango de los triglicéridos es más pequeño entonces a la hora de predecir, el modelo, ha tenido pocos diferentes a la media (todos rondan los mismos valores). Se crearon once modelos PLS y se seleccionaron los tres más relevantes.

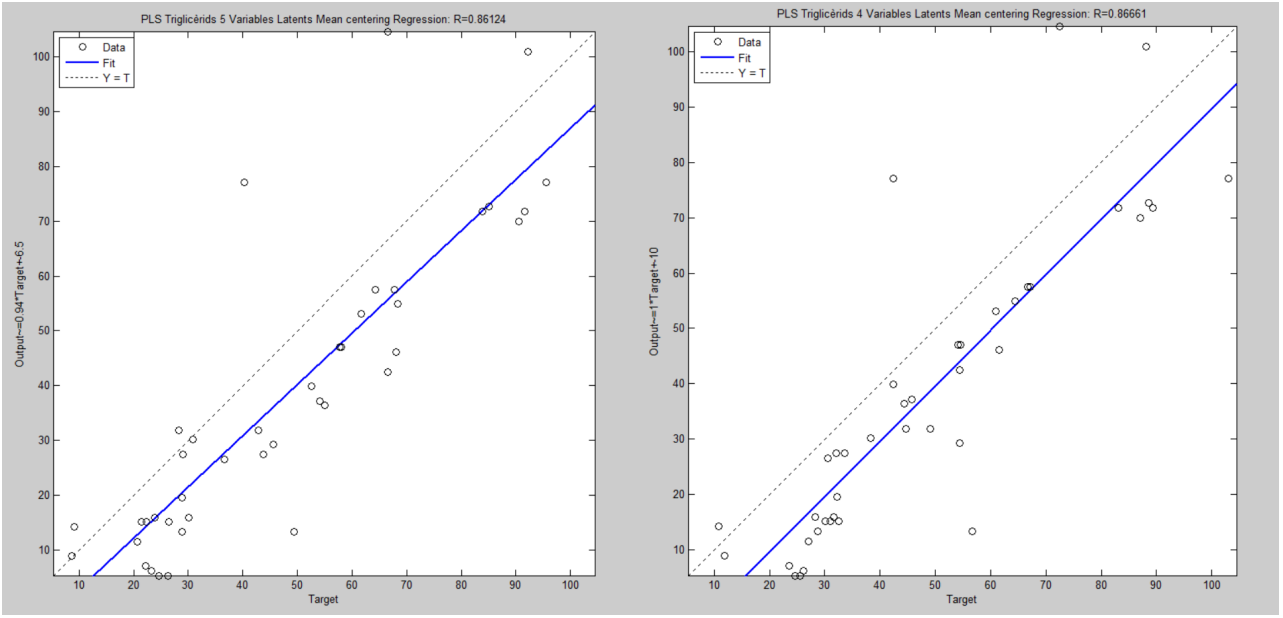


Figura 4-14. Modelo PLS para triglicéridos con 5 variables latentes, R = 0,86 y con preprocesado Mean centering (izquierda). Modelo PLS para triglicéridos con 4 variables latentes, R= 0,86661 y con preprocesado Mean centering (derecha).

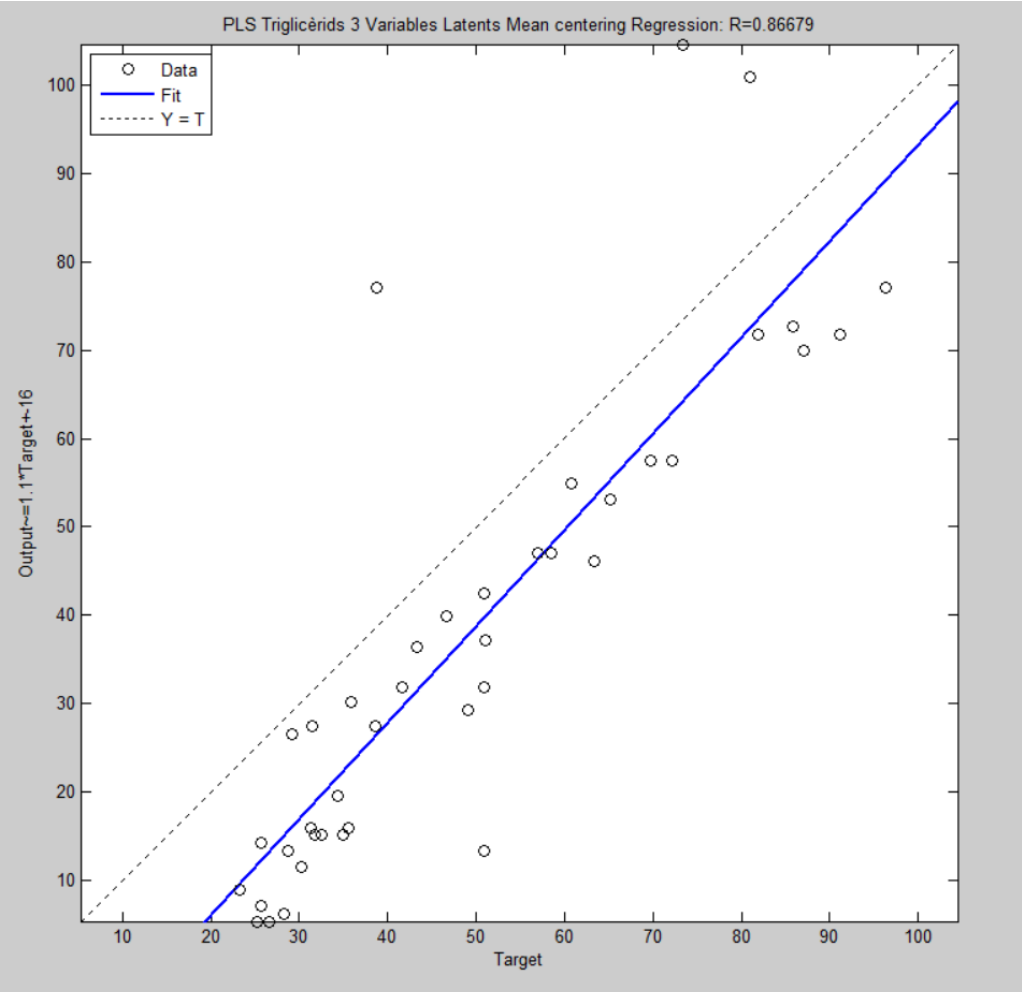


Figura 4-15. Modelo PLS para triglicéridos con 3 variables latentes, R = 0,86 y con preprocesado Mean centering.

Siguiendo los criterios tal y como se ha hecho en el caso anterior, se puede observar que el mejor modelo corresponde a la **Figura 4-15** ya que con una cantidad de variables latente menor consigue un coeficiente de correlación (*R*) del orden de 0,86 que también resulta ser el valor más elevado para triglicéridos de los 11 modelos PLS creados. A diferencia del caso anterior, el preprocesado Autoscale resulta en peores predicciones que el Mean centering. De hecho ninguno supera a estos tres para este caso.

Método 2 predicción de colesterol: Una vez finalizada la creación de modelos PLS para este primer método se procede a realizar lo mismo con el segundo método donde se garantizarán el 10% de los valores más elevados, 10% de los valores más bajos, el 50% de los valores aleatorios y el 30% restante para la prueba (**Figuras 4-16** y **4-17**). Este segundo método se aplica para garantizar el abanico de rangos más amplio posible ya que el entrenamiento abarcará desde los valores más grandes hasta los valores más pequeños.

Este método resultaría más interesante para la empresa ya que obligamos al modelo a entrenar con rangos de valores mucho más grandes, por lo que puede predecir una mayor variedad de muestras. El primer método no garantiza que los valores de los extremos (los más elevados y los más bajos) estén en el grupo de entrenamiento y por lo tanto no se puede saber cuán buenos resultados dará cuando tenga que predecir muestras con valores extremos. Para el colesterol con este método se crearon siete modelos PLS de los cuales se seleccionaron los tres mejores.

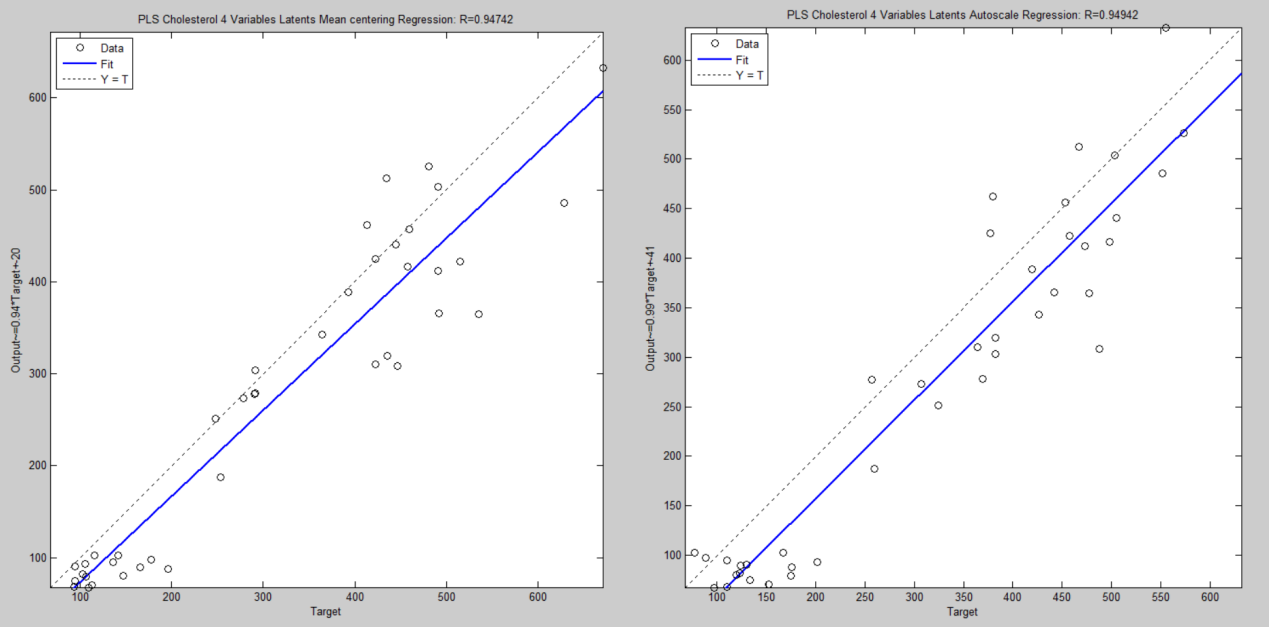


Figura 4-16. Modelo PLS para colesterol con 4 variables latentes, R = 0,94 y con preprocesado Mean centering (izquierda). Modelo PLS para colesterol con 4 variables latentes, R = 0,94 y con preprocesado Autoscale (derecha).

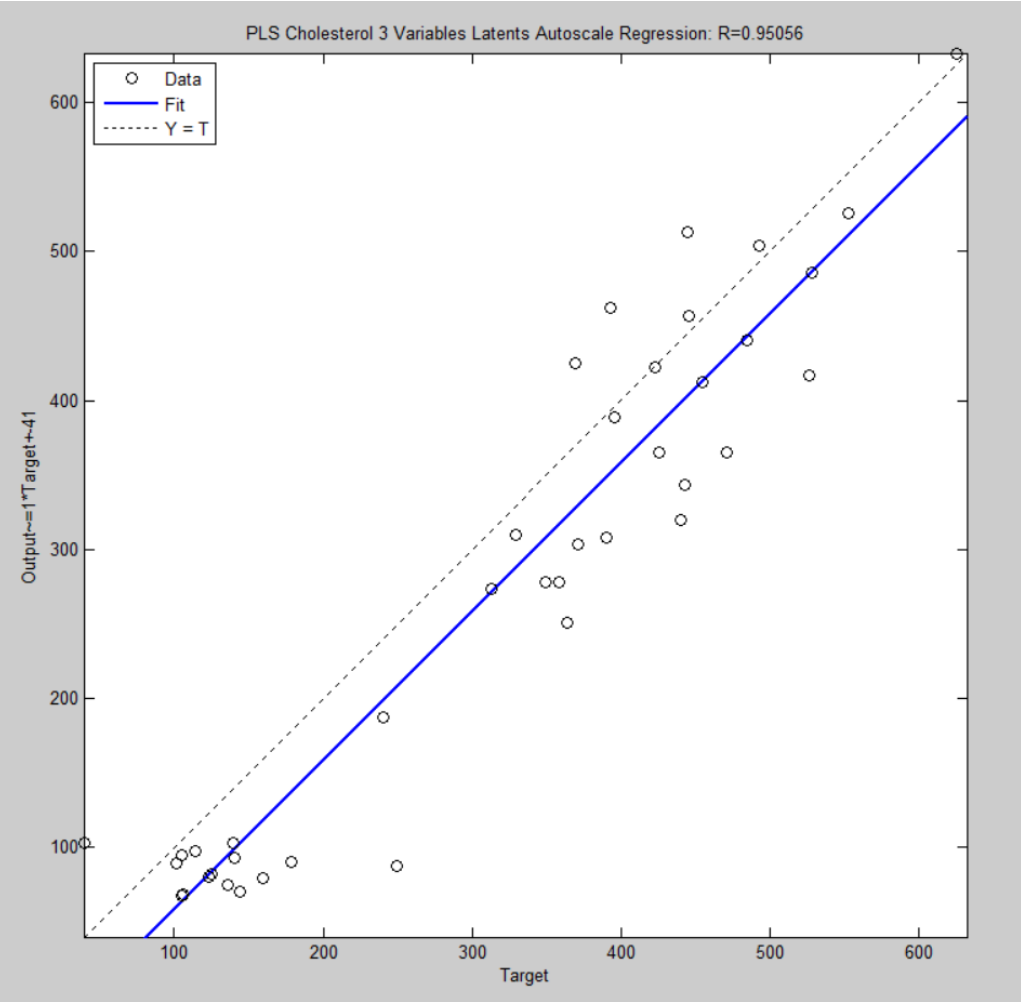


Figura 4-17. Modelo PLS para colesterol con 3 variables latentes, R = 0,95 y con preprocesado Autoscale.

Como se puede observar los resultados obtenidos tanto para el primer método como para el segundo método son similares en cuanto al colesterol, igual que la cantidad de variables latentes necesarias para conseguir dichos resultados. En el caso de la **Figura 4-17** resulta ser un modelo PLS más interesante que el mejor de los modelos para el primer método ya que este necesita 3 variables latentes para conseguir un valor de 0,95 mientras que el otro usa 4 variables latentes para un 0,97. Además, este método garantiza una buena predicción para valores extremos hecho que en el método 1 no se podía asegurar. Siguiendo la tendencia el preprocesado Autoscale garantiza un mejor resultado que el Mean centering.

Método 2 predicción de triglicéridos: Ahora hay que observar cómo se comporta el modelo PLS con el método 2 para los triglicéridos. Para ello se crearon siete modelos de los cuales se volvieron a seleccionar tres correspondientes a los mejores resultados (**Figuras 4-18** y **4-19**).

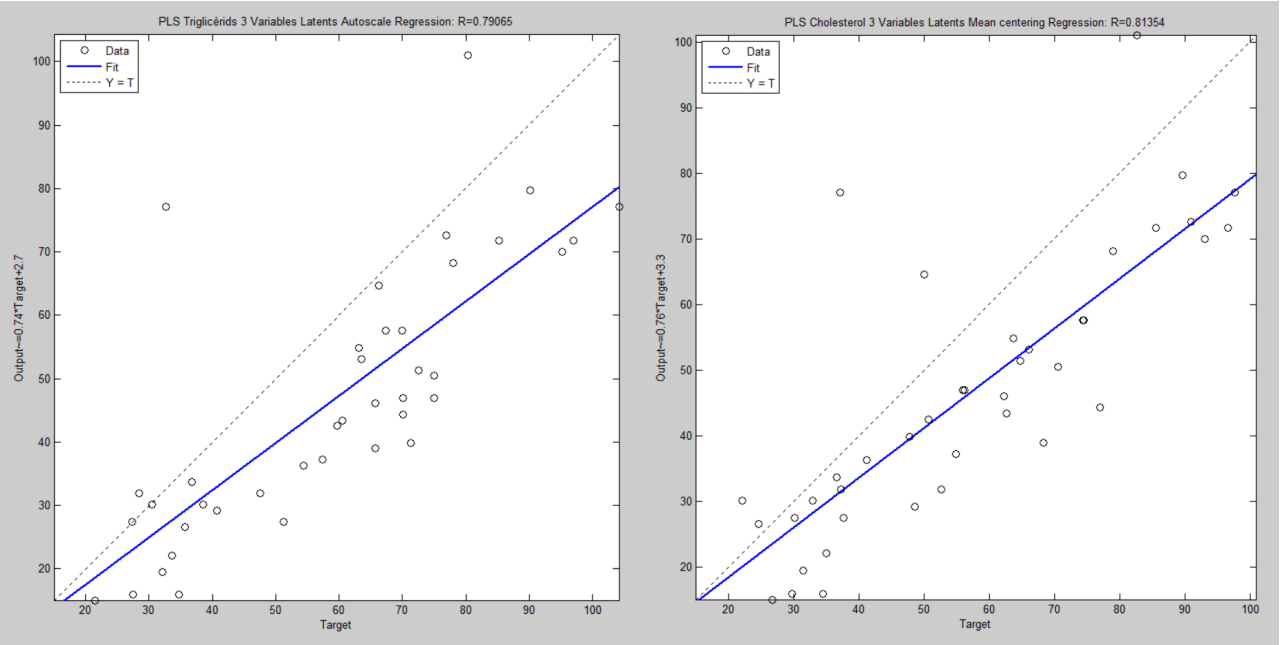


Figura 4-18. Modelo PLS para triglicéridos con 3 variables latentes, R = 0,79 y con preprocesado Autoscale (izquierda). Modelo PLS para triglicéridos con 3 variables latentes, R = 0,81 y con preprocesado Mean centering (derecha).

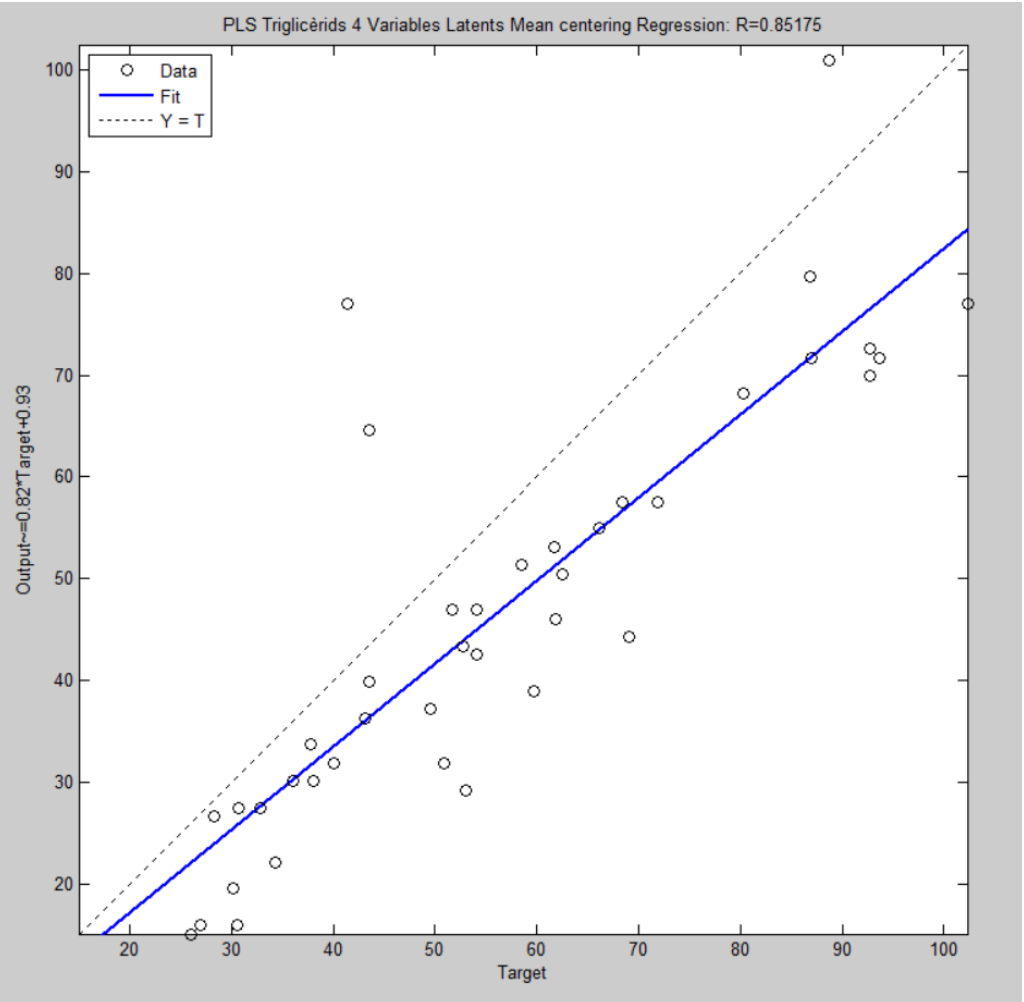


Figura 4-19. Modelo PLS para triglicéridos con 4 variables latentes, R = 0,85 y con preprocesado Mean centering.

Es evidente que con este método los triglicéridos se comportan notablemente peor que con el primero. No obstante, se ha logrado obtener un modelo que es capaz de predecir suficientemente bien los triglicéridos. De los siete modelos, el que se muestra en la **Figura 4-19** es el mejor dado que utiliza 4 variables latente y obtiene un coeficiente de correlación de 0,85 usando el preprocesado de Mean centering rompiendo con la tendencia en los otros modelos los cuales presentaban mejores resultados usando Autoscale como preprocesado.

De los cuatro modelos PLS más destacables (**Figuras 4-13, 4-15, 4-17** y **4-19**), dos corresponden al colesterol y dos corresponden a los triglicéridos. Para las dos lipoproteínas hay que seleccionar uno de los modelos como el más favorable para la empresa. Si se siguen los criterios de selección mencionados previamente y se considera que el método 2 garantiza un mayor rango de predicción, los mejores modelos corresponden a la **Figura 4-17** para el colesterol y a la **Figura 4-15** para los triglicéridos.

Una vez se han seleccionado los modelos PLS ya es posible disponer de un modelo capaz de cuantificar el colesterol y los triglicéridos totales, que era uno de los objetivos. A partir del código informático y del modelo de predicción seleccionado (uno para colesterol y otro para triglicéridos), cuando la empresa tenga nuevos conjuntos de datos de modelos animales podrá realizar la caracterización avanzada de lipoproteínas mediante la utilización de los modelos de predicción obtenidos en este trabajo de final de grado. No obstante se debe considerar que este estudio se ha realizado con dos de los tres conjuntos iniciales a causa de que el tercer conjunto no era comparable al de los otros dos, pero que es posible aumentar el número de conjuntos en cuanto la empresa obtenga más muestras y se analice las muestras.

4.5. Futuros trabajos

Uno de los nuevos proyectos que se podrían llevar a cabo a partir de este trabajo es la realización del mismo procedimiento con el fin de obtener modelos de predicción para otros tipos de modelos animales que sean de interés científico como es el caso de los ratones o las ratas.

Otro proyecto que no se ha realizado en este trabajo de final de grado a causa de la falta de tiempo y a la complejidad que engloba este tipo de proyectos sería usar los modelos obtenidos para crear un equivalente al test Liposcale® de la empresa en modelos animales de cerdo que les permita conseguir el perfilado avanzado de lipoproteínas y no solo la cuantificación de colesterol y triglicéridos totales.

5. Conclusiones

El presente trabajo ha desarrollado y evaluado diversos modelos de predicción PLS basados en la espectroscopia de resonancia magnética nuclear de protones. Dichos modelos han conseguido resultados prometedores que van a ser implementados en la compañía para la cuantificación de colesterol y triglicéridos en modelos animales de cerdo. Esto facilitará la labor de la empresa para poder obtener el perfilado avanzado de lipoproteínas basado en RMN en modelos animales a pesar de haber contado con tan solo dos conjuntos a la hora de realizar los modelos PLS, pero que se podrà aplicar a otros conjuntos siempre que estos resulten comparables con los usados en este proyecto.

Además de haber obtenido los modelos de predicción PLS, mediante el uso de Matlab, para poder cuantificar colesterol y triglicéridos totales, también se ha conseguido que sean modelos robustos y fiables con coeficientes de correlación más que satisfactorio. Como último objetivo era necesario saber distinguir que preprocesado ha resultado ser el más favorable, y tal como se puede ver en los resultados los modelos creados para predecir colesterol obtienen mejores resultados con el preprocesado Autoscale. Sin embargo, si intentamos crear modelos de predicción para triglicéridos, el mejor preprocesado es Mean centering. Como punto a destacar, es necesario recordar que los modelos se crearon para dos metodologías diferentes y que la segunda metodología, donde se procuraba garantizar el entrenamiento con valores extremos, resulta una opción más interesante que la primera metodología a la hora de obtener modelos de predicción PLS.

Para concluir con este trabajo, cabe mencionar que es posible crear un modelo PLS capaz de predecir fiablemente la concentración de colesterol y triglicéridos totales en muestras de plasma de cerdo, hecho que facilitará la labor de la empresa a la hora de obtener resultados analiticos y que podrán ser utilizados tanto en estudios de investigación como en ensayos clínicos para avanzar en el conocimiento y la prevención de las enfermedades cardiovasculares.

6. Referencias

1. World Health Organization. (n.d.). *Cardiovascular diseases*. World Health Organization. Retrieved June 6, 2022, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
2. Utermann, G. (1989). The mysteries of lipoprotein(a). *Science*, *246*(4932), 904–910. https://doi.org/10.1126/science.2530631
3. *Introduction to lipids and lipoproteins - NCBI bookshelf*. National Library of Medicine. (n.d.). Retrieved June 6, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK305896/
4. Judström-Kareinen, I. (2015). *Schematic drawing of lipoprotein structure. Information derived from Champe et al. 2005.* . ResearchGate. Retrieved June 14, 2022, from https://www.researchgate.net/figure/Schematic-drawing-of-lipoprotein-structure-Information-derived-from-Champe-et-al-2005_fig2_282356824.
5. *Introduction to lipids and lipoproteins - NCBI bookshelf*. National Library of Medicine. (n.d.). Retrieved June 6, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK305896/
6. P. Zipes MD, D. (2019). *Apolipoproteins*. Apolipoproteins - an overview | ScienceDirect Topics. Retrieved June 6, 2022, from https://www.sciencedirect.com/topics/neuroscience/apolipoproteins
7. Author links open overlay panelR WMahleyT LInnerarityS CRallJrK HWeisgraber, WMahley, R., LInnerarity, T., CRallJr, S., HWeisgraber, K., & Plasma lipoprotein metabolism is regulated and controlled by the specific apolipoprotein (apo-) constituents of the various lipoprotein classes. The major apolipoproteins include apoE. (1984, December 1). *Plasma lipoproteins: Apolipoprotein structure and function.* Journal of Lipid Research. Retrieved June 6, 2022, from https://reader.elsevier.com/reader/sd/pii/S0022227520344436?token=BF05F4140079F30CBF507DBB08A944D4C493CD7959F9C89528F71E9075B36155BA88E198477B01694A14A44CDA355288&originRegion=eu-west-1&originCreation=20220702114412
8. Elmadhun, N. Y., Sabe, A. A., Robich, M. P., Chu, L. M., Lassaletta, A. D., & Sellke, F. W. (2013). The pig as a valuable model for testing the effect of resveratrol to prevent cardiovascular disease. *Annals of the New York Academy of Sciences*, *1290*(1), 130–135. https://doi.org/10.1111/nyas.12216
9. Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., Richards, D. R., McDonald-Smith, G. P., Gao, H., Hennessy, L., Finnerty, C. C., López, C. M., Honari, S., Moore,

E. E., Minei, J. P., Cuschieri, J., Bankey, P. E., Johnson, J. L., Sperry, J., … Wong, W. H. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, *110*(9), 3507–3512. https://doi.org/10.1073/pnas.1222878110

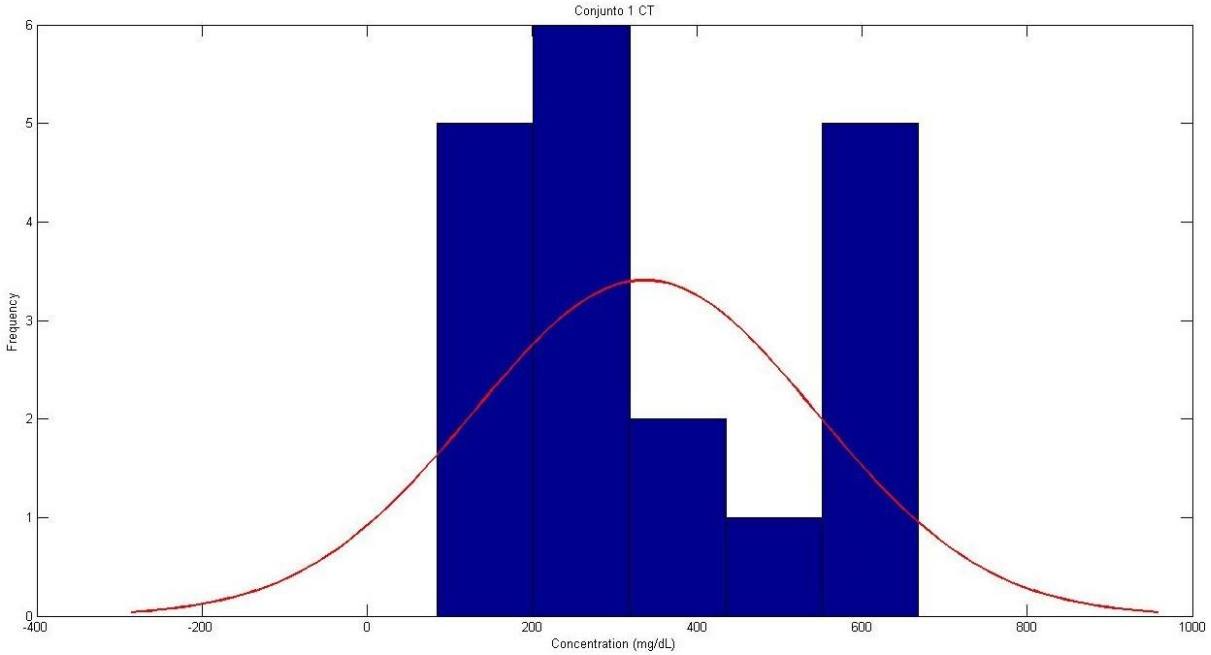
10. Yin, W., Carballo-Jane, E., McLaren, D. G., Mendoza, V. H., Gagen, K., Geoghagen, N. S., McNamara, L. A., Gorski, J. N., Eiermann, G. J., Petrov, A., Wolff, M., Tong, X., Wilsie, L. C., Akiyama, T. E., Chen, J., Thankappan, A., Xue, J., Ping, X., Andrews, G., … Strack, A. M. (2012, January). *Plasma lipid profiling across species for the identification of optimal animal models of human dyslipidemia*. Journal of lipid research. Retrieved June 9, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243481/
11. *Dyslipidemia - StatPearls - NCBI Bookshelf*. (n.d.). Retrieved June 9, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK560891/
12. Rauw, W. M., Portolés, O., Corella, D., Soler, J., Reixach, J., Tibau, J., Prat, J. M., Diaz, I., & Gómez-Raya, L. (2007). Behaviour influences cholesterol plasma levels in a pig model. *Animal*, *1*(6), 865–871. https://doi.org/10.1017/s1751731107000018
13. Francis, P. D. J., How to check for pulsus paradoxus? No Comments | Dec 31, About The Author Prof. Dr. Johnson Francis Former Professor of Cardiology, & Prof. Dr. Johnson Francis Former Professor of Cardiology. (n.d.). *All about cardiovascular system and disorders*. All About Cardiovascular System and Disorders. Retrieved June 9, 2022, from https://johnsonfrancis.org/professional/friedewald-equation-for-calculating-vldl-and-ldl/
14. *American Chemical Society*. (n.d.). Retrieved June 11, 2022, from https://pubs.acs.org/doi/full/10.1021/jacs.1c09637
15. *Department of Food Science*. Hanne Christine S. Bertram - Research - Aarhus University. (n.d.). Retrieved June 11, 2022, from https://pure.au.dk/portal/en/persons/hanne-christine-s-bertram(9ad2a927-38d2-4267-9382-a851bf3c7e06).html
16. Kulkarni, K. R., Garber, D. W., Marcovina, S. M., & Segrest, J. P. (1994). Quantification of cholesterol in all lipoprotein classes by the VAP-II method. *Journal of Lipid Research*, *35*(1), 159–168. https://doi.org/10.1016/s0022-2275(20)40123-3
17. *What is mass spectrometry?* Broad Institute. (2021, May 5). Retrieved June 14, 2022, from https://www.broadinstitute.org/technology-areas/what-mass-spectrometry
18. Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., & Wishart, D. S. (2019). NMR spectroscopy for Metabolomics Research. *Metabolites*, *9*(7), 123. https://doi.org/10.3390/metabo9070123

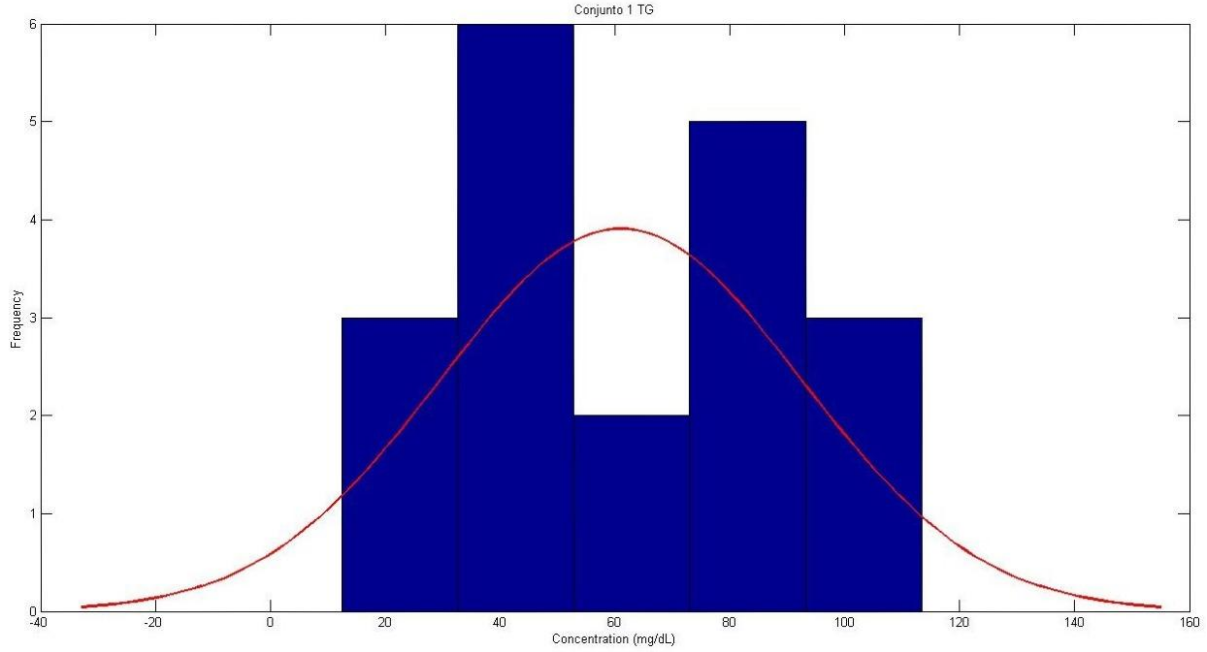
19. National Library of Medicine. (n.d.). *1H Nmr Spectrum of ML260*. National Library of Medicine. Retrieved June 14, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK143562/figure/ml260.f2/.
20. Mallol, R., Amigó, N., Rodríguez, M. A., Heras, M., Vinaixa, M., Plana, N., Rock, E., Ribalta, J., Yanes, O., Masana, L., & Correig, X. (2015). Liposcale: A novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. *Journal of Lipid Research*, *56*(3), 737–746. https://doi.org/10.1194/jlr.d050120
21. Valqui, P. by M., & Valqui, M. (2022, June 3). *NMR spectroscopy*. ChemTalk. Retrieved June 14, 2022, from https://chemistrytalk.org/nmr-spectroscopy/
22. *Colesterol Estructura de una molécula*. (n.d.). Dreamstime. Retrieved June 15, 2022, from https://es.dreamstime.com/colesterol-estructura-de-una-mol%C3%A9cula-image120252308.
23. *Ejemplo de un triglicérido graso insaturado*. (n.d.). Wikipedia. Retrieved June 15, 2022, from https://es.wikipedia.org/wiki/Triglic%C3%A9rido.
24. *Fórmula de un triglicérido formado con tres diferentes ácidos grasos*. (n.d.). Portal Académico CCH. Retrieved June 16, 2022, from https://e1.portalacademico.cch.unam.mx/alumno/quimica2/unidad2/grasas/trigliceridos.
25. Jeyarajah, E. J., Cromwell, W. C., & Otvos, J. D. (2006). Lipoprotein particle analysis by nuclear magnetic resonance spectroscopy. *Clinics in Laboratory Medicine*, *26*(4), 847–870. https://doi.org/10.1016/j.cll.2006.07.006
26. Mallol R, et al. Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. J Lipid Res (2015) NOU
27. Johnson Jr. CS Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. Progress in Nuclear Magnetic Resonance Spectroscopy 34(3-4):203-256 (1999).
28. Biosfer Teslab. (2021, October 4). Retrieved June 16, 2022, from https://biosferteslab.com/ca/qui-som/
29. Warmenhoven, J., Bargary, N., Liebl, D., Harrison, A., Robinson, M. A., Gunning, E., & Hooker, G. (2021). PCA of waveforms and functional PCA: A Primer for Biomechanics. *Journal of Biomechanics*, *116*, 110106. https://doi.org/10.1016/j.jbiomech.2020.110106
30. Beteinakis, S., Papachristodoulou, A., Gogou, G., Katsikis, S., Mikros, E., & Halabalaki, M. (2020). NMR-based metabolic profiling of edible olives-determination of quality parameters. *Molecules*, *25*(15). https://doi.org/10.3390/molecules25153339
31. Judström-Kareinen, I. (n.d.). *Schematic drawing of lipoprotein structure. Information derived from Champe et al. 2005.* . ResearchGate. Retrieved June 18, 2022, from

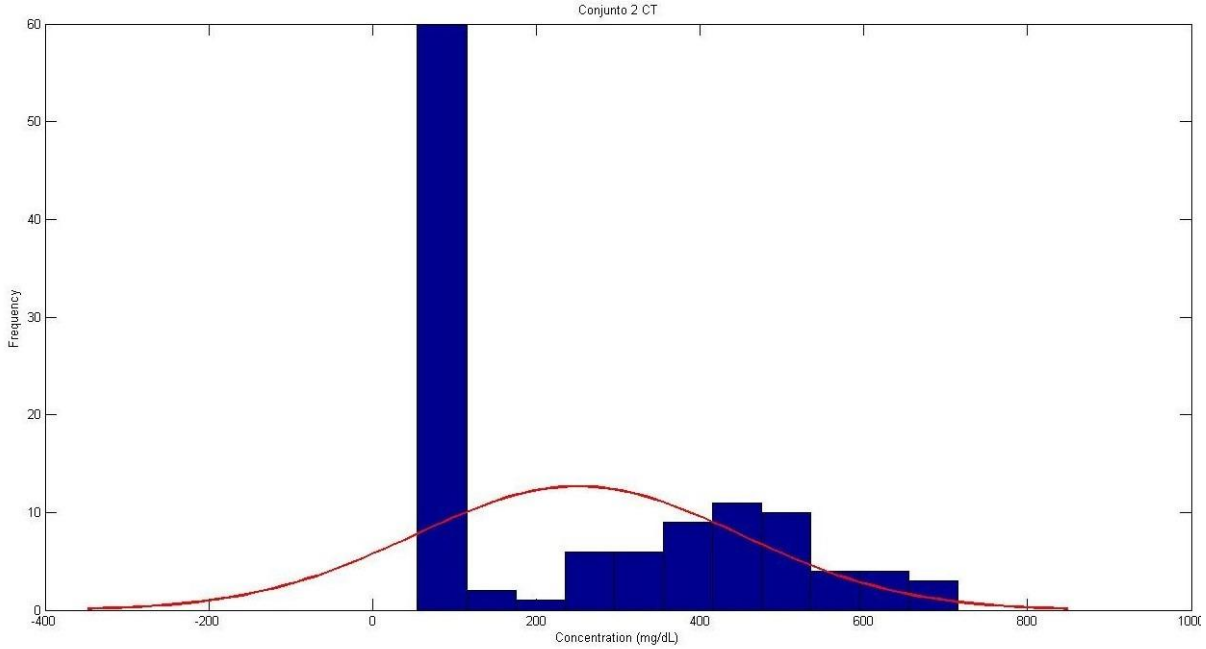
https://www.researchgate.net/figure/Schematic-drawing-of-lipoprotein-structure-Information-derived-from-Champe-et-al-2005_fig2_282356824.

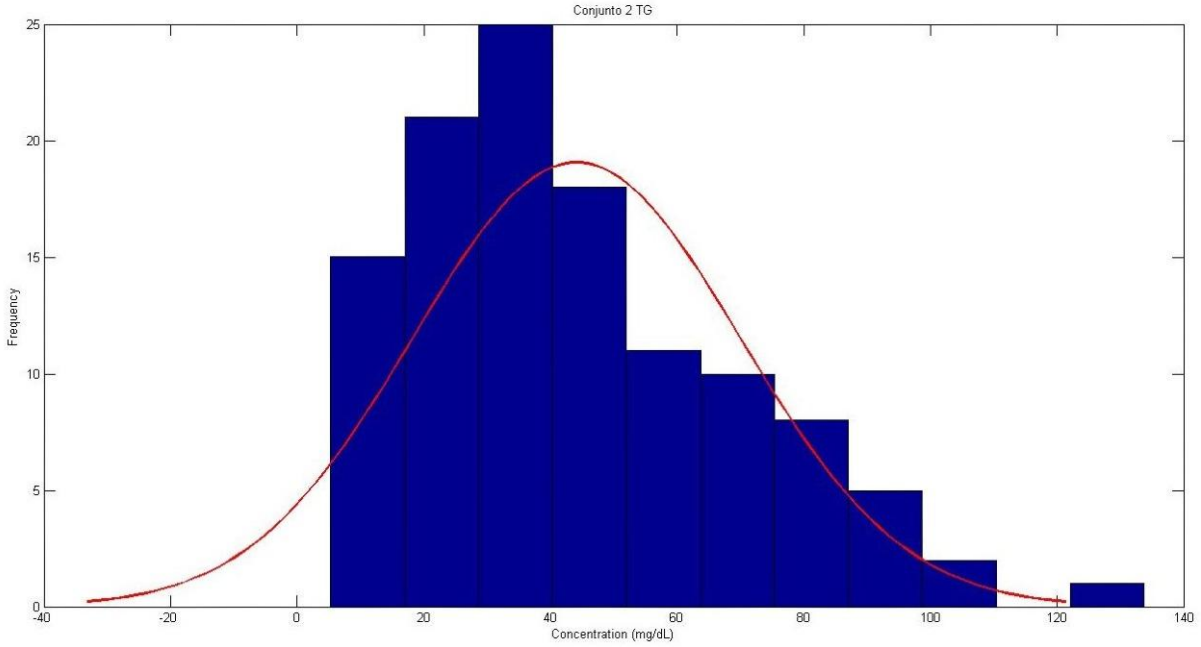
32. *Autoscale: Unit variance scaling method performed on the columns of the data (i.e. metabolite concentrations measured by 1H NMR or Binned 1H NMR spectra)*. RDocumentation. (n.d.). Retrieved June 18, 2022, from https://www.rdocumentation.org/packages/RFmarkerDetector/versions/1.0.1/topics/autoscale
33. Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *The Annals of Statistics*, *24*(2). https://doi.org/10.1214/aos/1032894467
34. Zargaran, A., Sakhteman, A., Faridi, P., Daneshamouz, S., Akbarizadeh, A. R., Borhani-Haghighi, A., & Mohagheghzadeh, A. (2017). Reformulation of traditional chamomile oil: Quality controls and fingerprint presentation based on cluster analysis of attenuated total reflectance–infrared spectral data. *Journal of Evidence-Based Complementary & Alternative Medicine*, *22*(4), 707–714. https://doi.org/10.1177/2156587217710982
35. Moreno,, R., Diego, J., Aguilar, M., & Didier, J. (2014, January). Niveles sanguíneos de colesterol y triglicéridos en cerdos alimentados con fruto entero de palma. Barrancabermeja-Colombia; Revista Citecsa.
36. *Cholesterol: Types, tests, treatments, prevention*. Cleveland Clinic. (n.d.). Retrieved June 13, 2022, from https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean
37. MediLexicon International. (n.d.). *Cholesterol levels by age: Health Ranges, what is high, and tips*. Medical News Today. Retrieved June 13, 2022, from https://www.medicalnewstoday.com/articles/315900#levels-and-age
38. *The truth about triglycerides*. The Truth About Triglycerides - Health Encyclopedia - University of Rochester Medical Center. (n.d.). Retrieved June 13, 2022, from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=56&contentid=2967
39. *Triglycerides: Health risks, ways to lower levels*. Cleveland Clinic. (n.d.). Retrieved June 13, 2022, from https://my.clevelandclinic.org/health/articles/11117-triglycerides

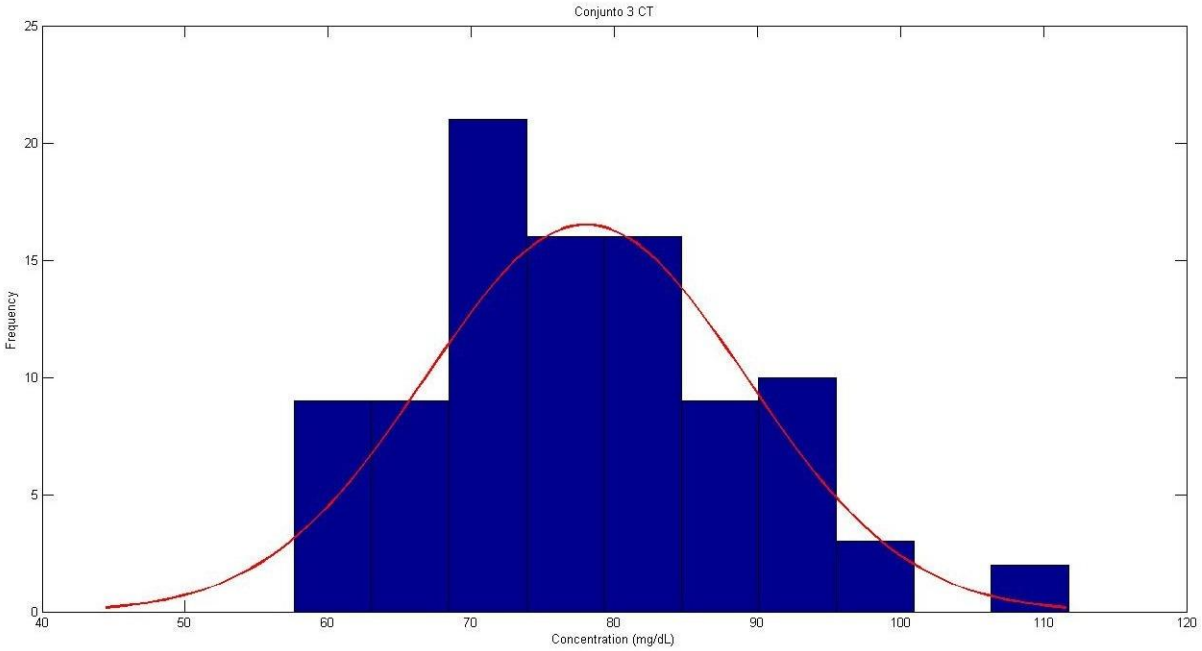
Anexo I. Histogramas

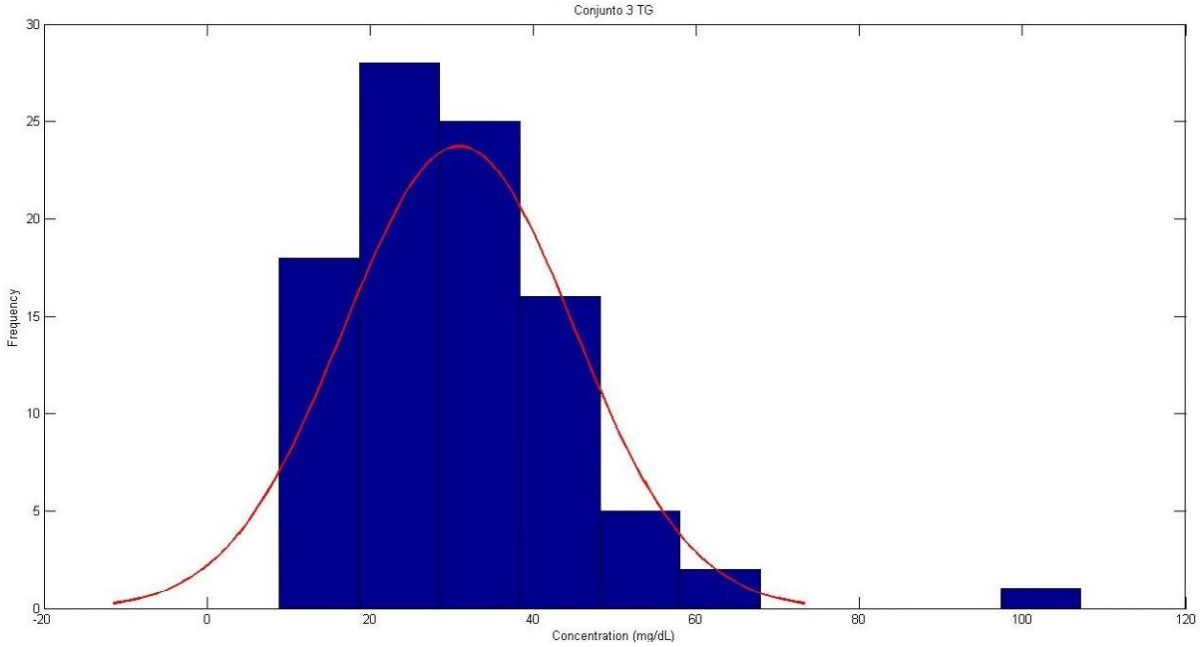





Conjunto 2 CT
Frequency
Concentration (mg/dL)
60
50
40
30
20
10
0
-400
-200
0
200
400
600
800
1000


Conjunto 2 TG
Frequency
Concentration (mg/dL)
25
20
15
10
5
0
-40
-20
0
20
40
60
80
100
120
140


Conjunto 3 CT
Frequency
Concentration (mg/dL)
0
5
10
15
20
25
40
50
60
70
80
90
100
110
120


Conjunto 3 TG
Frequency
Concentration (mg/dL)
0
5
10
15
20
25
30
-20
0
20
40
60
80
100
120

Anexo II. Esquema del proceso