Búsqueda de endofenotipos de enfermedades respiratorias mediante la aplicación de técnicas de agrupamiento no supervisado

Universitat Oberta de Catalunya

Álvaro García Muñoz

Análisis multivariante de datos ómicos

Bioinformática y Bioestadística UOC-UB

Nombre del tutor del TFM:
Dr. Jose Luis Mosquera Mayo
(Externo) José Miguel Lorenzo Salazar

Nombre del PRA:
Dr. Carles Ventura

18 de junio de 2024

UNIVERSITAT DE BARCELONA

Ficha Del Trabajo Final

Título del trabajo:	Búsqueda de endofenotipos de enfermedades respiratorias mediante la aplicación de técnicas de agrupamiento no supervisado
Nombre del autor/a:	Álvaro García Muñoz
Nombre del tutor del TFM:	Dr. Jose Luis Mosquera Mayo (Externo) MSc. José M. Lorenzo-Salazar
Nombre del PRA:	Dr. Carles Ventura
Fecha de entrega:	18 de junio de 2024
Titulación o programa:	Bioinformática y Bioestadística UOC-UB
Área del trabajo final:	Análisis multivariante de datos ómicos
Idioma del trabajo:	Castellano
Palabras clave:	clustering, supervised, unsupervised, endophenotypes, hierarchichal clustering, multivariate analyisis, neuronal network, topological

Resumen del trabajo

Un endofenotipo es un rasgo biológico o de comportamiento medible que está genéticamente relacionado con una enfermedad. Los endofenotipos son considerados biomarcadores. Los biomarcadores juegan un rol importante porque desarrollan los métodos de diagnóstico, prevención y detección de enfermedades.

Este TFM plantea como hipótesis la existencia de endofenotipos que permitan encontrar biomarcadores en enfermedades respiratorias, específicamente EPID y COVID-19 grave tratado con corticoides. Para la validación de la hipótesis se realiza un análisis aplicando métodos de agrupamiento no supervisado como K-means y HDBSCAN, desarrollando a su vez una interfaz de usuario capaz de aplicar esto métodos de agrupamiento no supervisado. Se ha conseguido detectar pequeños agrupamientos densos de pacientes de EPID y muchos agrupamientos de pacientes de COVID-19 grave tratado con corticoides.

En el futuro, será necesario seguir indagando y contrastando los resultados aquí obtenidos, continuar la investigación aumentando la potencia estadística en los datos EPID y realizar un análisis genético de los agrupamientos encontrados en los datos de COVID-19. También se sugiere ampliar la interfaz de usuario a otros métodos de agrupamiento no supervisado, implementando técnicas de aprendizaje profundo y técnicas de análisis topológico de datos.

Abstract

An endophenotype is a measurable biological or behavioural trait that is genetically related to a disease. Endophenotypes are considered biomarkers. Biomarkers play an important role in developing methods of diagnosis, prevention and detection of diseases.

This TFM hypothesises the existence of endophenotypes that allow biomarkers to be found in respiratory diseases, specifically DPLD and severe COVID-19 treated with corticosteroids. In order to validate the hypothesis, an analysis is carried out by applying unsupervised clustering methods such as K-means and HDBSCAN, while developing a user interface capable of applying these unsupervised clustering methods. Small dense clusters of DPLD patients and many clusters of severe COVID-19 patients treated with corticosteroids have been detected.

In the future, it will be necessary to further investigate and contrast the results obtained here, to continue the research by increasing the statistical power in the DPLD data and to perform a genetic analysis of the clusters found in the COVID-19 data. It is also suggested to extend the user interface to other unsupervised clustering methods, implementing deep learning techniques and topological data analysis techniques.

Agradecimientos

El desarrollo de este Trabajo de Fin de Máster ha sido posible gracias al apoyo inestimable del equipo del grupo de investigación del Área de Genómica del Instituto Tecnológico y de Energías Renovables (ITER), en particular mi tutor externo José Miguel Lorenzo Salazar, por su invaluable guía. Este trabajo tampoco hubiera sido posible sin la ayuda de mi tutor de la UOC, Dr. José Luis Mosquera Mayo, por su seguimiento y valiosas sugerencias. A todos, mi más sincero agradecimiento.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo.

La pandemia causada por el SARS-CoV-2, agente etiológico de la COVID-19, ha puesto de manifiesto la relevancia de las enfermedades respiratorias como un problema de salud pública de gran magnitud [2]. Ello ha incrementado significativamente la investigación en el campo de las enfermedades infecciosas respiratorias, impulsando el desarrollo de nuevas estrategias de prevención, diagnóstico y tratamiento [3].

En este contexto, el presente Trabajo de Fin de Máster (en adelante TFM) constituye una investigación basada en el estudio de datos ómicos de alto rendimiento orientado a la búsqueda de endofenotipos. Los endofenotipos son las distintas formas que tiene de manifestarse una enfermedad según el paciente que la padece. En concreto, en este trabajo se investigan las Enfermedades Pulmonares Intersticiales Difusas o EPIDs y la COVID-19 grave.

Para la búsqueda de endofenotipos se aplican métodos dirigidos a la detección de posibles biomarcadores. Con este objetivo, se ha desarrollado un flujo bioinformático programado en R con técnicas de aprendizaje automático (*Machine Learning*). Estos algoritmos permiten la identificación no supervisada de agrupamientos en datos de proteómica de alto rendimiento procedentes de pacientes con EPIDs, y datos genéticos de pacientes de COVID-19 grave tratados con corticoides durante su hospitaliación.

La búsqueda de endofenotipos es un proceso clave para la aplicación y el desarrollo médico de tratamientos [4]. La identificación de características que permitan clasificar un fenotipo determinado en varias clases o endofenotipos [5] facilita una mayor precisión en el diagnóstico, contribuye a la detección temprana de las enfermedades, permite una selección más específica de tratamientos. Además, fomenta la búsqueda de biomarcadores que orienten la creación de nuevos medicamentos o la reutilización de los existentes. Todo esto va dirigido a posibilitar una mejor comprensión de la enfermedad y desarrollar tratamientos más eficaces en la lucha contra dicha enfermedad [6] .

Para la búsqueda e identificación de endofenotipos será necesario aplicar el análisis multivariante, ya que permite realizar agrupamientos de individuos a partir de observaciones clínicas,

de imagen, o moleculares, entre otras. En términos matemáticos, este análisis requiere de la definición de espacios geométricos métricos, en los que mediante la introducción de la distancia entre los individuos se realice el agrupamiento. Estas distancias permiten establecer relaciones de equivalencia entre los datos (o pacientes) gracias a los algoritmos de agrupamiento no supervisado [7].

Estos algoritmos han visto un crecimiento exponencial debido al desarrollo de las técnicas de aprendizaje automático y los recursos informáticos, en particular aquellos relacionados con el denominado *Big Data*. La existencia de numerosos métodos de agrupamiento no supervisado se asocia con la posibilidad de encontrar varios resultados diferentes según los parámetros escogidos. Es por ello que, entre los objetivos del trabajo, se ha incluido un aspecto centrado en la replicación y reproducibilidad de la investigación.

1.2. Objetivos del trabajo.

Este trabajo tiene como objetivo general implementar algoritmos de agrupamiento no supervisado en datos reales de enfermedades respiratorias. Para cumplir con este objetivo se han planteado tres objetivos principales, que dan lugar a seis objetivos secundarios, que se exponen a continuación:

1.2.1. Objetivos generales.

1. Estudiar estrategias de agrupamiento supervisado y no supervisado para la clasificación de pacientes que presentan enfermedades respiratorias con fenotipos y endofenotipos, conocidos y desconocidos.

2. Desplegar un flujo de trabajo bioinformático y una interfaz de usuario [8] que permita el agrupamiento no supervisado para el descubrimiento de los endofenotipos desconocidos de enfermedades respiratorias.

3. Aplicar el flujo de trabajo bioinformático a datos reales de pacientes con enfermedades respiratorias, como son las EPIDs y la COVID-19 grave.

1.2.2. Objetivos específicos.

- 1.a. Realizar un análisis bibliográfico de técnicas de agrupamiento (*clustering*), supervisado y no supervisado, basado tanto en aprendizaje automático (*machine learning*) como en aprendizaje profundo (*deep learning*).

- 1.b. Realizar un análisis bibliográfico de métodos de reducción de dimensiones y de métricas orientadas a la comparación de utilidad en biomedicina.

- 2.a. Realizar pruebas de *clustering* utilizando distintos enfoques (geométrico y topológico) en matrices de datos genéticos.

- 2.b. Crear un flujo de trabajo bioinformático y una interfaz de usuario para la búsqueda de endofenotipos en matrices de datos genéticos de carácter masivo.
- 3.a. Aplicar el flujo de trabajo bioinformático y la interfaz de usuario programada en shiny en enfermedades respiratorias.
- 3.b. Analizar los resultados de la aplicación del flujo de trabajo bioinformático y la interfaz de usuario programada en enfermedades respiratorias, poniendo en relación los resultados y la definición matemática del algoritmo.

1.3. Impacto de sostenibilidad, ético-social y de diversidad.

En el contexto del trabajo, el impacto de sostenibilidad se refiere a la capacidad del estudio para generar resultados que tengan un impacto duradero. Esto se ve reflejado en el desarrollo de las herramientas para hacer accesible la aplicación de métodos de agrupamiento no supervisado.

Los métodos de agrupamiento no supervisado aplicados en datos obtenidos de pacientes de enfermedades neurodegenerativas, como el Alzheimer [9], han demostrado ser herramientas potentes para la detección de estructuras subyacentes que permiten la detección de biomarcadores.

Las estructuras y patrones subyacentes observados en los datos obtenidos para una determinada enfermedad permiten la detección de endofenotipos, impulsando de esta manera el desarrollo de la llamada *medicina de precisión* o *medicina personalizada*. Es decir, aquella medicina que busca *"maximizar la calidad de la asistencia sanitaria individualizando el proceso de atención sanitaria en función de la evolución única del paciente"* [10].

Recientemente, gracias al desarrollo de las herramientas de obtención de proteomas humanos y el desarrollo de la investigación de estos datos, se han logrado encontrar biomarcadores de alto rendimiento diagnóstico [11]. En este contexto, el aprendizaje automático mejorará significativamente el campo del desarrollo de herramientas proteómicas [12].

Sin embargo, se ha de tener en cuenta las consideraciones éticas y sociales al realizar el estudio y comunicar los resultados. Esto implica proteger la confidencialidad de los datos comunicando los resultados de manera clara, precisa y responsable, teniendo en cuenta las implicaciones para los pacientes y profesionales de la salud.

Las EPIDs y, particularmente la FPI o Fibrosis Pulmonar Idiopática, son enfermedades pulmonares graves que presentan grandes problemas a la hora de ser diagnosticadas [13]. Una de las mayores implicaciones que se encuentran es la posibilidad de encontrar biomarcadores que faciliten el diagnóstico de estas enfermedades respiratorias.

No es la primera ocasión en la que se aplican técnicas de aprendizaje automático sobre datos de proteómica [14], pero sí resulta innovador el enfoque dado en este trabajo en la aplicación

de métodos de agrupamiento no supervisado que permitan la detección de endofenotipos [15] sobre datos de proteómica de alto rendimiento.

Reflejar el impacto de diversidad es involucrar en la muestra, la totalidad de la población estudiada. Por ello, en el trabajo realizado se estudia una muestra representativa de la población en términos de edad, sexo y ascendencia de la cual se poseen datos.

1.4. Enfoque y método seguido.

Tal y como se ha expuesto previamente en el contexto del apartado 1.1, este trabajo se desarrolla adoptando un enfoque matemático y bioinformático. El trabajo se ha nutrido de la definición de espacios métricos dentro de los conceptos dotados por la inferencia estadística y la topología.

Los espacios métricos proporcionan un marco formal para cuantificar la distancia entre puntos en un conjunto. En la inferencia estadística, los espacios métricos se utilizan para definir conceptos como la convergencia de secuencias de variables aleatorias, la consistencia de estimadores y la normalidad asintótica. Estos conceptos son esenciales para analizar datos y evaluar la precisión de los métodos estadísticos.

La metodología seguida para desarrollar el trabajo se basa en el Método Científico. El Método Científico se desarrolla en diferentes etapas, desde la definición del problema, la formulación de hipótesis y los objetivos de trabajo, como la propia metodología experimental, la obtención de resultados, la discusión y evaluación de los mismos, la obtención de conclusiones y la difusión del trabajo. Cada una de estas etapas se desarrolla en los distintos capítulos de la presente memoria.

A continuación, se detallan las etapas seguidas en el trabajo:

- En la parte inicial del TFM, se ha recopilado una importante cantidad de bibliografía para entender bien el problema de base en la identificación de endofenotipos en el contexto de la aplicación de las tecnologías ómicas de alto rendimiento en determinadas enfermedades complejas, como pueden ser las enfermedades neurodegenerativas [11, 16], cardiovasculares [17], cáncer [4], etc. En la última década, gracias al desarrollo de la proteómica de alto rendimiento, los estudios de estas enfermedades han progresado considerablemente, permitiendo la identificación de determinados biomarcadores que facilitarán la medicina de precisión y el tratamiento de múltiples enfermedades.
- A la vista del contexto definido, se ha planteado como hipótesis la siguiente afirmación: En el conjunto de datos ómicos de alta dimensión obtenidos con técnicas de alto rendimiento, de genotipado mediante array o secuenciación masiva, y de proteómica dirigida de alto rendimiento, existen endofenotipos no identificados en pacientes de enfermedades respiratorias graves, como la fibrosis pulmonar idiopática o la COVID-19 grave.
- Con el objetivo de buscar e identificar endofenotipos en los datos proporcionados para desarrollar el TFM, se adopta un enfoque multivariante usando métodos de agrupamiento no supervisado sobre distintos conjuntos de datos ómicos.

- En el diseño de la experimentación se planificó y desarrolló la generación de un informe en formato RMarkDown. Este informe dinámico, fácil de ejecutar, es capaz de llevar a cabo el análisis multivariante de este tipo de datos pudiendo replicarse con datos ajustados de la misma forma. Gracias a las funciones diseñadas en el citado informe dinámico, se ha desarrollado una interfaz de usuario en shiny que permite el agrupamiento no supervisado de los datos y su visualización con estadísticos que permiten su elección en función de los objetivos propuestos.
- Una vez realizado el diseño experimental, y preparadas las herramientas bioinformáticas, se han aplicado a dos conjuntos de datos separados. Se han aplicado dichas herramientas en datos de proteómica de alto rendimiento de un estudio de pacientes afectados por EPIDs a y también se han aplicado en datos genotípicos de un estudio de mortalidad de pacientes afectados por la COVID-19 grave tratados con corticoides.
- Finalmente se han analizado los datos y se ha extraído una serie de conclusiones que se detallan en el **Capítulo 4** de este documento.

1.5. Planificación del trabajo.

Para llevar a cabo el trabajo se han identificado varias herramientas informáticas (Google Drive; GitHub; OverLeaf) y se han definido tareas concretas que permitieran el control del desarrollo del mismo. Las tareas se han agrupado en función de los objetivos y cada una de ellas se ha concretado para ser realizada en un tiempo definido conforme al cronograma previsto.

- Tareas del objetivo 1.a:
  - A1.2 - Comprender los algoritmos por los que se rigen los diferentes métodos de reducción de dimensiones y agrupamiento supervisado o no supervisado.
  - A1.3 - Identificar, instalar y probar diversos paquetes de software libre de R y Python diseñados para realizar clustering y detectar agrupamientos empleando técnicas diversas de análisis supervisado y no supervisado.
- Tareas del objetivo 1.b:
  - A1.1 - Realizar una búsqueda bibliográfica en PubMed definiendo palabras clave relacionadas con el objeto del TFM (IPF, ILD, Unsupervised Clustering, Supervised Clustering, Endophenotype, Biomarker, Multivariate Analysis, Clustering Methods...).
- Tareas del objetivo 2.a:
  - A2.1 - Crear un entorno de trabajo en RStudio que facilite el acceso a los programas.
  - A2.2 - Definir e instalar los paquetes necesarios para el desarrollo de la investigación del TFM en todas sus vertientes.
- Tareas del objetivo 2.b:

- A2.3 - Estudiar los formatos de archivo de variación genética: archivo de genotipos generado tras procesar microarrays (archivos ped/map, bed/bim/fam) y archivos procedentes de secuenciación masiva (VCF). Aprender a gestionar (leer, manipular, convertir, etc.) las matrices de datos de variación genética, comprender su uso y entender qué representan.
- A2.4 - Realización pruebas iniciales de clustering supervisado y no supervisado para la familiarización con el entorno informático que permita el uso de datos en distintos formatos de entrada de demostración (datos denominados *toy* o datos para pruebas).
- A2.5 - Realización de un análisis comparativo de los resultados en función del algoritmo de clustering utilizado y el método de reducción de dimensiones empleado para la visualización gráfica.

- Tareas del objetivo 3.a:
  - A3.1 - Desarrollo de diferentes scripts en R que permitan la gestión de los datos, análisis multivariante y la aplicación de técnicas de agrupamiento no supervisado.
  - A3.2 - Desarrollo de un informe dinámico en formato RMarkDown que implemente el código desarrollado en los scripts para el análisis de datos reales de proteómica de alto rendimiento y datos fenotípicos de los estudios descritos en apartados anteriores.
  - A3.3 - Desarrollo de una interfaz de usuario en shiny [8] que facilite la implementación y visualización de métodos de agrupamiento no supervisado.

- Tareas del objetivo 3.b:
  - A4.1 - Aplicación del informe en RMarkDown y la interfaz de usuario en shiny a los datos reales procedentes de los estudios de alto rendimiento de EPIDs y de COVID-19 grave.
  - A4.2 - Análisis y discusión de los resultados obtenidos. Detallados en el capítulo 3.

A continuación se muestra la temporalización de las tareas mediante un diagrama de Gantt 1.1. En él se detalla la previsión de los tiempos para realizar las diferentes tareas. En el diagrama distinguimos dos grupos de tareas, aquellas propuestas por la UOC (en azul) y las tareas propias del trabajo (en naranja).

Una vez se han definido los objetivos y tareas que permiten la evaluación del desarrollo del trabajo, se procede a explicar el impacto de esta investigación.

1.6. Breve sumario de productos obtenidos

Este trabajo ha generado varios productos tangibles. Estos productos se encuentran detallados en el Capítulo 2. En resumen, son:

- Un repositorio público de GitHub.
- Un informe dinámico desarrollado en RMarkDown.
- Una interfaz de usuario elaborada en R usando el paquete shiny.

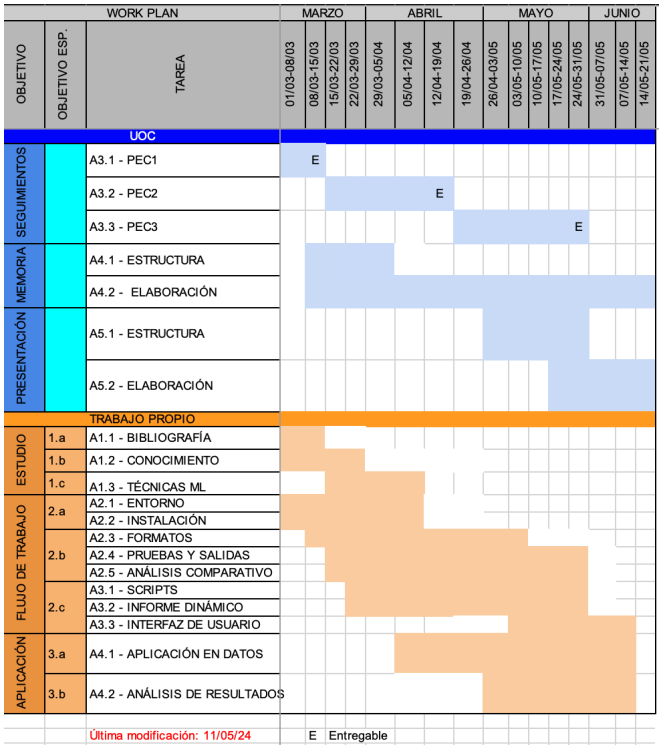

Figura 1.1: Diagrama de Gantt con la temporalización de las distintas tareas planificadas en el TFM.

1.7. Breve descripción de los otros capítulos de la memoria

Este documento de memoria del TFM se ha dividido en cuatro capítulos.

El **Capítulo 1** permite situar el trabajo en su contexto, su impacto y los objetivos propuestos para el desarrollo del mismo.

El **Capítulo 2** explica los datos analizados en el trabajo de investigación, los métodos de análisis implementados y los productos obtenidos durante el desarrollo.

El **Capítulo 3** expone los resultados de la aplicación de los métodos de análisis en los datos reales.

Por último, el **Capítulo 4** presenta las conclusiones del trabajo desarrollado, los limites encontrados y los posibles futuros trabajos.

Capítulo 2

Materiales y métodos.

A lo largo de este capítulo se describen las herramientas empleadas para el desarrollo de los objetivos de las tareas previstas. Se describen los materiales y recursos utilizados en la investigación incluyendo los conjuntos de datos, el software estadístico y los recursos computacionales. También se detallan los pasos seguidos para la recopilación y preprocesamiento de los datos y la selección de los algoritmos de agrupamiento y de reducción de dimensiones.

2.1. Formato y procedencia de los datos

Los datos biológicos, cuando se preparan para su tratamiento informático, se pueden utilizar de diferentes formas. Estos formatos van a depender de la naturaleza de la información a conservar. Los dos formatos principales son PLINK y VCF.

A. PLINK

En el caso de los datos obtenidos mediante el genotipado del ADN de pacientes utilizando la tecnología de microarrays, es frecuente trabajar con el formato conocido como PLINK. En este caso, se genotipa un número conocido de posiciones del genoma (generalmente entre 0,5 y 1 millón de posiciones del genoma). En este formato se utilizan distintas combinaciones de archivos en función de la información que albergan: podemos encontrarnos con los datos guardados en formato ped/map (o en su formato binario bed/bim/fam). Los archivos ped (o su versión binaria, bed) almacenan la información de genotipado por líneas. Los archivos map (o su alternativa, el tándem bim/fam) contienen información sobre la posición de cada variante, su identificación y los individuos analizados.

B. VCF

Cuando se trabaja con datos procedentes de la secuenciación masiva de ácidos nucleicos generalmente se utiliza el formato VCF (Variant Calling Format). Este formato está pensado para describir la variación genética observada en uno o más individuos (en cuyo caso se habla

de VCF multimuestra) así como sus anotaciones. El archivo VCF canónico contiene una cabecera y un cuerpo con los datos de la variación genética descrita. En el cuerpo del archivo VCF se presenta la información sobre la localización de cada variante (cromosoma y posición), los alelos que constituyen la variante (alelo de referencia y alelo alternativo), una serie de campos de anotación propios del nombrado de la variante (esto es, del procedimiento de identificación de la variación utilizando herramientas bioinformáticas específicas) y el genotipo, además de múltiples campos de anotación biológica, entre otros. Toda esta información se organiza por columnas en el cuerpo del archivo VCF (que contrasta con el esquema interno de los archivos en formato PLINK).

En el proceso de investigación se ha trabajado con estos dos formatos principales, PLINK y VCF.

Los datos empleados en el desarrollo del trabajo se clasifican en dos grupos diferenciados en función de su formato y uso:

- **Datos de demostración.** Se trata de datos orientados a facilitar el desarrollo y aprendizaje del uso del software y de los diferentes entornos de trabajo. A saber:
  - Datos genéticos en **formato VCF** extraídos del cromosoma 22 de 715 individuos no relacionados de siete poblaciones diferentes del Proyecto 1000 Genomas [18]. Se han utilizado para estudiar el funcionamiento y aplicación de los métodos de agrupamiento supervisado y no supervisado, así como para los métodos de reducción dimensional presentados en los apartados precedentes. No se dispone de variables adicionales (covariables). El fenotipo o variable *target* es la población declarada en el portal de datos del Proyecto 1000 Genomas.
  - Datos genéticos en **formato PLINK** del proyecto **Servicio de Análisis de Datos Genómicos (SAMDG)** desarrollado en el área de Genómica del ITER, y disponibles a través de su página web: SAMDG [19]. En este caso, no se dispone de variables adicionales (covariables). El fenotipo o variable target es la población declarada en los propios datos ofrecidos en el portal web de ITER.
- **Datos reales:** Datos procedentes de la aplicación de técnicas ómicas en los dos estudios de enfermedades respiratorias graves que completan el desarrollo del trabajo. Estos datos provienen de dos investigaciones en curso actualmente y han sido publicados parcialmente:
  - Datos de **EPIDs:** Datos de proteómica de alto rendimiento obtenidos con tecnología PEA desarrollada por Olink en pacientes con EPID debidamente anonimizados. Se han estudiado un total de 369 proteínas en 432 pacientes. Se desarrollan en el **apartado 2.1.1** de la memoria. También se dispone de la siguiente información: sexo, edad, centro de reclutamiento y ascendencia (autodeclarada). La variable fenotípica o *target* es el tipo de EPID diagnosticada.
  - Datos de **COVID-19 grave en pacientes tratados con corticoides:** Datos obtenidos en el marco del consorcio nacional SCOURGE [20] y han sido anonimizados. Se han obtenido las diez primeras componentes principales relativas a los datos genéticos de los pacientes del estudio que han sido tratados con corticoides. Se desarrollan

en el **apartado 2.1.2**. También se dispone de la siguiente información: sexo, edad y centro de reclutamiento. La variable fenotípica o *target* es la mortalidad a 90 días (todos los pacientes han sido tratados con corticoides).

2.1.1. Datos de proteómica de alto rendimiento.

Entre las técnicas para la obtención de datos de proteómica más utilizados se encuentran la Cromatografía Líquida de Alto Rendimiento acoplada con Espectrometría de Masas (HPLC-MS/MS) y la Espectrometría de Proteínas por Electroforesis de Afinidad (PEA).

La MS o MS/MS (cuando se acoplan dos detectores de masa) es una técnica analítica que permite identificar y caracterizar moléculas mediante la medición de su relación masa/carga. En el contexto de la proteómica, la MS se utiliza para identificar y cuantificar proteínas mediante la **fragmentación de sus péptidos** constituyentes tras pasar por un proceso de separación inicial a través de un equipo de cromatografía líquida de alta resolución (HPLC).

La PEA es una técnica que se basa en la **separación de proteínas por electroforesis** en un gel de agarosa, seguido de su inmovilización en la superficie del gel y su **detección mediante anticuerpos específicos**. Esta técnica permite obtener un perfil proteico de la muestra analizada, proporcionando información sobre la presencia, abundancia y modificaciones postraduccionales de las proteínas. En un contexto de alto rendimiento, nos encontramos con la tecnología PEA desarrollada por la empresa **Olink**. Esta empresa ha desarrollado varios tipos de paneles con hasta 384 proteínas objetivo. Estos paneles permiten la realización de inmunoensayos multiplex de alto rendimiento utilizando volúmenes mínimos de suero, plasma o casi cualquier otro tipo de muestra biológica. Generalmente, se requiere de sistemas de automatización para conseguir resultados reproducibles y trazables. Estos, a su vez, requieren de equipos de secuenciación masiva para obtener los perfiles proteómicos finales de cada individuo. Por tanto, se trata de una tecnología híbrida que acopla la tecnología PEA propietaria de Olink con la secuenciación masiva (NGS, de *Next Generation Sequencing*) de ácidos nucleicos.

Los datos obtenidos se expresan en formato **NPX**, por sus siglas en inglés *Normalized Protein eXpression*. NPX es una unidad arbitraria establecida por Olink en la que se escalan los datos utilizando un logaritmo de base 2 (muy conveniente porque un cambio de una unidad implica que la expresión de la proteína se duplica). Se calcula a partir del valor de Ct (ciclo umbral de una PCR) y se realiza con el objetivo de minimizar la variación tanto dentro del ensayo, como entre ensayos. Para el cálculo del valor se sigue el diagrama de flujo de la **Figura 2.1**.

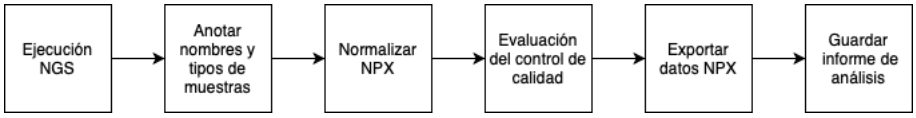

Figura 2.1: Diagrama de flujo del pre-procesamiento de los datos en la plataforma de proteómica de alto rendimiento de Olink [1].

En la **Figura 2.1** se muestra el preprocesamiento de los datos desde su obtención mediante

la combinación de PEA+NGS, hasta la normalización de los datos siguiendo el formato NPX (adaptado de la web de Olink).

En cuanto al análisis de dos proteínas mediante los valores de NPX, es importante saber que estas no se pueden comparar directamente. Los datos que se obtienen son datos de concentración de proteínas medidos en pg/ml. Sin embargo, tras el tratamiento de los datos y su transformación en datos NPX, los valores de concentración originales se desvirtúan debido a la calibración del modelo en función de los datos originales. Esto provoca que según la calibración realizada para cada proteína, un valor $X\ NPX$ pueda representar diferentes concentraciones. Por esta razón, se recomienda usar los valores NPX para la comparación de individuos y no para la comparación de proteínas.

Los valores NPX se calculan según la siguiente expresión:

$$NPX_{i,j} = ExtNPX_{i,j} - Med(ExtNPX(Controls_i)) \tag{2.1}$$

donde i se refiere a una proteína concreta, j se refiere a una muestra concreta, $Controls_i$ hace referencia al valor obtenido en las placas de control, la aplicación $Med()$ calcula la mediana y $ExtNPX$ define una aplicación de la extensión normalizada de un valor NPX, de la siguiente manera:

$$ExtNPX_{i,j} = \frac{log_2(contar(muestra_j, proteina_i))}{contar(ExtCt_j)} \tag{2.2}$$

donde la aplicación $contar()$ cuenta los objetos y el la variable $ExtCt$ es el valor de la cantidad generalizado.

Los datos de proteómica de alto rendimiento se han obtenido para un total de 369 proteínas (un panel) en 432 individuos no relacionados de un estudio en curso sobre EPID en el que participa el ITER. Como se ha señalado previamente, y dado que los resultados no han sido publicados todavía, el estudio es ciego para el estudiante, de forma que tanto las proteínas como los individuos son anónimos.

Además de los datos de proteómica de alto rendimiendo, de cada individuo se proporcionan datos de las siguientes covariables: sexo, edad, centro de reclutamiento, ascendencia y diagnóstico principal (**Cuadro 2.1**).

Entre las covariables destaca la variable de diagnóstico principal, que será utilizada como variable objetivo en los análisis realizados con este conjunto de datos en la investigación. Aunque el diagnóstico también se ha utilizado de forma ciega, esto es, desconocida por el estudiante, en el caso de las EPIDs podemos encontrarnos con dos grupos principales de fenotipos. Por un lado, se presentan síndromes heterogéneos que requieren una evaluación de las características clínicas, radiográficas y patológicas [21] para poder emitir un diagnóstico certero. Entre los síndromes más frecuentes, destaca la denominada enfermedad del tejido conectivo (ETC, o CTD en inglés). Para diagnosticar la ETC deben seguirse los criterios del Colegio Americano de Reumatología (ACR) [13]. Por otro lado, otro de los síndromes asociados a las EPIDs es la FPI, y para

Covariables	Datos del estudio	
Sexo	Hombre	$n = 289 (66{,}9\,\%)$
	Mujer	$n = 143 (33{,}1\,\%)$
Edad (años)	$min = 22$	
	$max = 100$	
	$media = 70{,}4$	
	$mediana = 72$	
Ancestralidad (autodeclarada)	$A =$ Asiático	$n = 20 (4{,}63\,\%)$
	$B =$ Negro	$n = 35 (8{,}10\,\%)$
	$H =$ Hispano	$n = 20 (4{,}63\,\%)$
	$O =$ Otra	$n = 2 (0{,}463\,\%)$
	$W =$ Blanco	$n = 355 (82{,}3\,\%)$
Diagnóstico principal	Diagnóstico 1	$n = 305 (70{,}6\,\%)$
	Diagnóstico 2	$n = 40 (9{,}26\,\%)$
	Diagnóstico 3	$n = 29 (6{,}71\,\%)$
	Diagnóstico 4	$n = 58 (13{,}42\,\%)$

Cuadro 2.1: Sumario de las covariables y fenotipos del estudio de proteómica de alto rendimiento en enfermedad pulmonar intersticial.

su diagnóstico se requiere la exclusión de enfermedades autoinmunes u otras causas. Ambos síndromes pueden dar lugar a fibrosis del parénquima pulmonar y, además, pueden compartir un patrón de neumonía intersticial habitual (NIU) en el análisis por imagen y biopsia. Algunos casos de ETC pueden distinguirse con relativa facilidad de la FPI. Sin embargo, en casos más complejos se requiere de la participación de un equipo multidisciplinar para consensuar un diagnóstico probable. Por tanto, nos encontramos con una patología con síntomas que pueden ser compatibles con enfermedades pulmonares distintas y con un diagnóstico complejo. En el caso del conjunto de datos de proteómica de alto rendimiento, el diagnóstico proporcionado se presenta en cuatro grupos diferenciados.

2.1.2. Datos de COVID-19 grave de pacientes tratados con corticoides.

Para la aplicación de métodos no supervisados en datos reales, se ha usado también un segundo conjunto de datos. Se trata de un estudio en curso sobre la mortalidad a 90 días utilizando datos genéticos de 15.000 pacientes de COVID-19 de España, entre los que algunos han sido hospitalizados y han sido tratados con corticoides. Estos datos provienen del estudio SCOURGE [22], en el que se reclutaron casos de COVID-19 en 34 centros de 25 ciudades del territorio nacional.

Una vez recogida la muestra de sangre periférica, se realizó la extracción del material genético y se realizó el genotipado utilizando el microarray Axiom Spain Biobank Array (Thermo Fisher Scientific) [22]. Tras la realización de múltiples controles de calidad al uso, se finaliza con un archivo de genotipos estándar en formato ped/map. A partir de este archivo (no disponible

para el estudiante), se realizó un PCA, facilitando tanto el archivo generado por PLINK para los autovectores propios (componentes principales) como los autovalores (útiles para generar el *screeplot*).

En este estudio se proporcionan otras variables, además de las componentes principales obtenidas del archivo de genotipos de los pacientes (**Cuadro 2.2**). A saber: sexo, edad y mortalidad. Dado el interés en la respuesta de los pacientes graves al tratamiento con corticoides y que no existe definición actual de lo que implica una mala respuesta a dicho tratamiento, se prestó interés a la posible identificación de endofenotipos relacionados con la mortalidad.Para ello se utilizó como variable *target* la mortalidad a los 90 días.

Covariables	**Datos del estudio**	
Sexo	Hombre	$n = 941 (63{,}3\,\%)$
	Mujer	$n = 545 (36{,}7\,\%)$
Edad (años)	$min = 19$	
	$max = 102$	
	$media = 71{,}91$	
	$mediana = 73$	
Hospital	Centro 1	$n = 81 (5{,}45\,\%)$
	Centro 2	$n = 89 (5{,}99\,\%)$
	Centro 3	$n = 92 (6{,}19\,\%)$
	Centro 4	$n = 50 (3{,}36\,\%)$
	Centro 5	$n = 24 (1{,}62\,\%)$
	Centro 6	$n = 295 (19{,}85\,\%)$
	Centro 7	$n = 48 (3{,}23\,\%)$
	Centro 8	$n = 792 (53{,}3\,\%)$
	Centro 9	$n = 15 (1{,}01\,\%)$
Mortalidad (90 días)	Sobreviven	$n = 1198 (80{,}5\,\%)$
	No sobreviven	$n = 288 (19{,}4\,\%)$
	Datos faltantes	$n = 1 (0{,}06\,\%)$

Cuadro 2.2: Sumario de las covariables y fenotipos del estudio de proteómica de alto rendimiento en enfermedad pulmonar intersticial.

2.2. Flujo de trabajo bioinformático.

El flujo de trabajo bioinformático (*pipeline*) diseñado para desarrollar los objetivos 2 y 3 del trabajo se expone a continuación (**Figura 2.2**), partiendo de los datos de entrada y finalizando en la generación de gráficos de salida con los agrupamientos encontrados.

El flujo de trabajo bioinformático se inicia con el paso de lectura de los datos bioinformáticos en el formato adecuado (archivos en formato PLINK, VCF, tabular, etc.). A continuación, tras comprobar que la lectura es correcta (individuos en filas) y que se han identificado los nombres

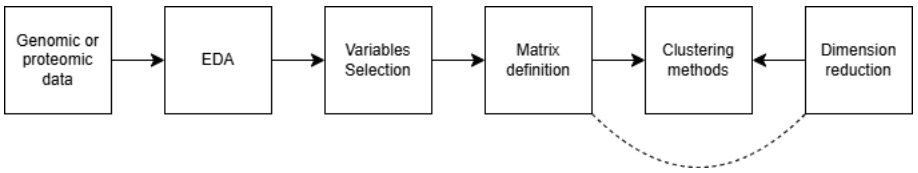

Figura 2.2: Diagrama del flujo de trabajo bioinformático diseñado en el TFM.

de cada variable (en columnas), se realiza un Análisis Exploratorio de los Datos (*Exploratory Data Analysis* o EDA).

El EDA se realiza tanto de forma univariante (cada variable por separado),como bivariante (variables de entrada tomadas de dos en dos; por ejemplo, se estudia la expresión semicuantitativa por parejas de proteínas para determinar su nivel de correlación), obteniendo también las distribuciones empíricas de cada variable. El EDA también comprende un estudio multivariante para detectar la presencia de posibles valores anómalos o *outliers* en el estudio. No obstante, dado que el objetivo es la búsqueda de endofenotipos a través del uso de métodos no supervisados, no se eliminarán estos valores anómalos, sino que quedarán marcados para ser visualizados en el resto de los análisis.

El flujo de trabajo bioinformático continúa a partir del EDA implementando una selección de variables. Es posible seleccionar todas las variables proporcionadas para el estudio o se pueden usar algoritmos como RFE (de sus siglas en inglés *Recursive Feature Elimination*), que permitan aplicar algún criterio de elección de variables para reducir las dimensiones del espacio de entrada de los datos. Por ejemplo, es esperable encontrar dos o más proteínas que pertenezcan a la misma ruta metabólica cuyos valores de expresión semicuantitativa estén correlacionados. La cuestión aquí radica en decidir si una de estas dos variables debe ser eliminada y, llegado el caso, en decidir cuál de ellas debe ser eliminada, toda vez que el estudio es ciego (ver **Apartado 4.2** de limitaciones y trabajos futuros).

Una vez realizado el apartado de selección de variables, se procede a la definición de matrices. Para el manejo de los datos estudiados en el trabajo se han definido 3 matrices complementarias. Las matrices se encuentran definidas en la **Figura 2.3**.

La **Figura 2.3** muestra la estructura de las matrices definidas para el análisis de los datos. Las filas de la matriz son comunes y se corresponden a los individuos de cada estudio. Las variables se presentan en columnas. En la figura, X_a representa las variables biológicas de los individuos (genómicas o proteómicas), mientras que las variables Y_b y Z representan las variables que acompañan a los datos biológicos, siendo Z el fenotipo o variable target que se desea estudiar. Las matrices definidas son las indicadas en la parte inferior de la figura: una matriz de datos biológicos, una matriz con las variables estudiadas, biológicas y no biológicas (sin incluir el fenotipo), y una matriz extendida que incluye todos los datos seleccionados para el estudio.

Tras definir las matrices de datos, se han aplicado los métodos de reducción de dimensiones para la visualización de los resultados explicados en el **Apartado 2.3**, y los métodos de

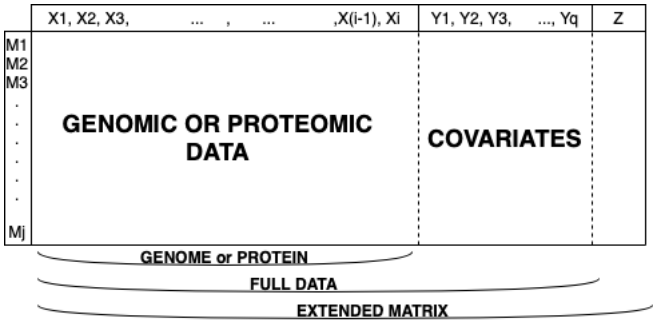

Figura 2.3: Estructura de las diferentes matrices.

agrupamiento no supervisado explicados en el **Apartado 2.4**.

2.3. Métodos de reducción de dimensiones.

Tal y como hemos indicado en el capítulo 1, el trabajo que se desarrolla con las proteínas requiere del análisis multidimensional. Es por ello, que en el ámbito del análisis de datos se utilizan con frecuencia estrategias para la reducción de dimensiones y lograr, de esta forma, un estudio más comprensible de los datos con los que contamos. La reducción de dimensiones se refiere a un conjunto de herramientas y algoritmos que permiten transformar un conjunto de datos de mayor dimensión en un subconjunto de menor dimensión sin pérdida de información relevante. En el ámbito bioestadístico, al estudiarse muestras de gran cantidad de datos (secuencias genéticas o proteínas), los métodos de reducción de dimensiones buscan reducir la complejidad de muestras representadas por múltiples variables. Estos métodos son útiles para diversas aplicaciones, como la reducción de ruido, entendiendo por ruido una variabilidad inexplicable dentro de una muestra de datos, o la visualización de los datos pues pueden ser proyectados en un espacio de dimensión 2D o 3D. En el desarrollo de este trabajo, las técnicas seleccionadas para la reducción de dimensiones permitirán principalmente la visualización de los datos. Se han seleccionado los siguientes métodos:

- **Análisis de componentes principales (PCA):** Es un método lineal enfocado en la transformación de variables. Este método transforma las variables originales en un nuevo conjunto de variables no correlacionadas (las componentes principales), ordenadas por su varianza decreciente [23].
- **Escalado multidimensional (MDS):** Es un método que permite conservar las distancias entre los puntos de datos en el espacio original al mapearlos en un espacio de menor dimensión.
- **Incrustación Estocástica de Vecinos distribuidos en t (t-SNE):** Es un método no lineal enfocado en la incrustación estocástica de vecinos distribuidos en t, según su probabilidad usando la distribución t de Student [24]. Los resultados pueden variar con diferentes inicializaciones y parámetros, como el número de pasos de la iteración y el valor

del parámetro *perplexity*. Esta técnica no es adecuada para la reducción de dimensionalidad orientada al aprendizaje automático, sino más bien para la visualización.

- **Aproximación y Proyección Uniforme de Variedades (UMAP):** Es un método no lineal enfocado en el análisis topológico de los datos y los conjuntos de complejos simpliciales [25]. El algoritmo de fondo es diferente a t-SNE, aunque sus usos son similares a los descritos para dicha técnica.

A continuación, vamos a profundizar en el estudio de cada uno de estos métodos y su aportación al desarrollo de este trabajo, para la visualización de resultados se utilizaran los datos de prueba explicados en el **apartado 2.1**

2.3.1. Análisis de Componentes Principales (PCA).

El Análisis de Componentes Principales o PCA (de *Principal Component Analysis*) es un método algebráico clásico de reducción de dimensiones. El objetivo principal del método es la construcción de nuevas componentes ortogonales sobre las que se proyectan los objetos de alta dimensión originales. Dado que las nuevas variables no presentan correlación entre ellas, se evita la redundancia estadística en el análisis y los posibles problemas asociados a la colinealidad. Este comportamiento se debe a una importante propiedad de la covarianza. Sean A y B dos variables aleatorias:

$$cov(A, B) = cov(B, A) \tag{2.3}$$

Si suponemos $x_1, \cdots, x_n$ una muestra de variables aleatorias, su matriz de varianzas y covarianzas es la siguiente:

$$\begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & \ddots & & \vdots \\ \vdots & & \ddots & cov(x_{n-1}, x_n) \\ cov(x_n, x_1) & \cdots & cov(x_n, x_{n-1}) & cov(x_n, x_n) \end{pmatrix} \tag{2.4}$$

Por lo tanto, la matriz 2.4 es una matriz cuadrada de tamaño $n \times n$ simétrica por la propiedad 2.3. Esto demuestra que la matriz de varianzas y covarianzas es diagonalizable.

Los valores propios de la matriz $y_1, ..., y_n$ definirán la varianza de unas nuevas variables o componentes principales independientes, tras la definición de la matriz del mismo tamaño en la base definida por los vectores propios de esta misma. Al ordenarse los valores propios de forma que $y_i \geq y_j \iff i > j$, permite que las nuevas variables o componentes principales explican la varianza de las muestras ordenadas de mayor a menor.

En la **Figura 2.4** se observa el diagrama denominado *scree-plot*, un gráfico que muestra el porcentaje de varianza explicada por cada componente principal tras la aplicación del método PCA a datos genéticos provenientes de un conjunto de donantes no relacionados tomados de

tres superpoblaciones (EUR, SAS y AFR) y siete poblaciones (CEU, FIN, GBR, IBS, TSI, CHB y YRI). Estos datos se han tomado del *Proyecto 1000 Genomas* en dos realizaciones distintas (para 50 y 100 componentes principales). El análisis de las representaciones (**Figura 2.4**) muestra que las dos primeras componentes principales (PC1 y PC2) representan una parte importante de la varianza del conjunto original de datos.

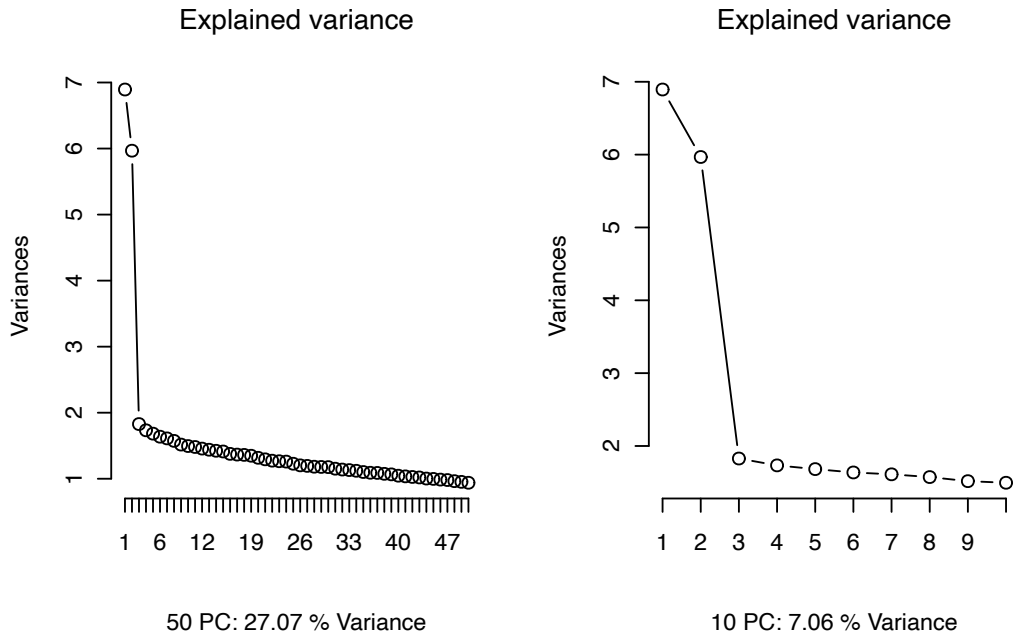

Figura 2.4: *Scree-plot* de la aplicación del PCA sobre la información genética procedente de 715 individuos no relacionados procedentes de tres superpoblaciones (EUR, SAS y AFR) y siete poblaciones (CEU, FIN, GBR, IBS, TSI, CHB y YRI) para dos realizaciones distintas (50 PCs, izquierda; 100 PCs, derecha).

Esto nos permite identificar las primeras dos componentes principales, PC1 y PC2, como aquellas que nos permitirán representar el conjunto de datos original, para lo que bastará utilizar un espacio bidimensional. Dicho espacio describe de forma sucinta la estructura poblacional que resulta de las frecuencias alélicas en las poblaciones estudiadas (**Figura 2.5**), tal y como ha sido descrito [26].

2.3.2. Escalado Multidimensional (MDS).

El Escalado Multidimensional o MDS (de *Multidimensional Scaling*) es un método no lineal de reducción de dimensionalidad. Se basa en la conservación de las distancias entre los puntos en alta dimensión al representarlos en baja dimensión, asumiendo que las relaciones entre los puntos pueden aproximarse mediante distancias métricas, es decir este algoritmo busca conservar las distancias calculadas entre puntos en alta dimensión.

El MDS se basa en la minimización de una función de pérdida que mide la diferencia entre las distancias originales entre los puntos en el espacio de alta dimensión y las distancias entre los

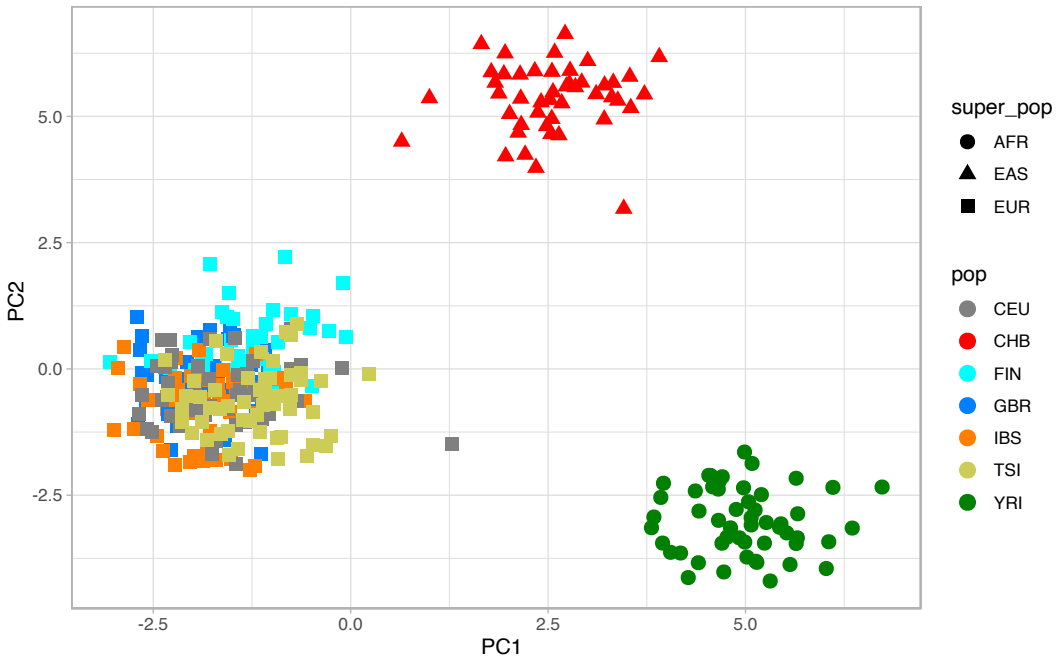

Figura 2.5: Representación de la varianza genética con las dos primeras componentes principales tras la aplicación de la PCA sobre 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas.

puntos proyectados en baja dimensión. En el MDS clásico se utiliza la tensión de Kruskal [27], definida a continuación, como función de pérdida:

$$S = \sum_{i,j} w_{ij} \|d_{ij} - d'_{ij}\|^2 \tag{2.5}$$

donde S representa la tensión de Kruskal, w_{ij} es un peso para medir la importancia de las distancia entre los puntos i y j, d_{ij} es la distancia original en el espacio de alta dimensión y d'_{ij} es la distancia en el espacio de baja dimensión.

La función de pérdida se minimiza iterativamente utilizando un algoritmo de optimización como, por ejemplo, el descenso de gradiente. Así, en cada iteración, se actualizan las posiciones de los puntos en el espacio de baja dimensión para minimizar la tensión de Kruskal actualizando las posiciones de los puntos de la siguiente forma:

$$y_i = y_i - \alpha \times \Delta_S(y_i) \tag{2.6}$$

donde α es el paso de aprendizaje y $\Delta_S(y_i)$ es el gradiente de tensión con respecto a la posición del punto i. Este gradiente se calcula a partir de la expresión 2.5.

En la **Figura 2.6** se muestra en análisis de MDS aplicado sobre los datos del estudio presentado el **apartado 2.1**.

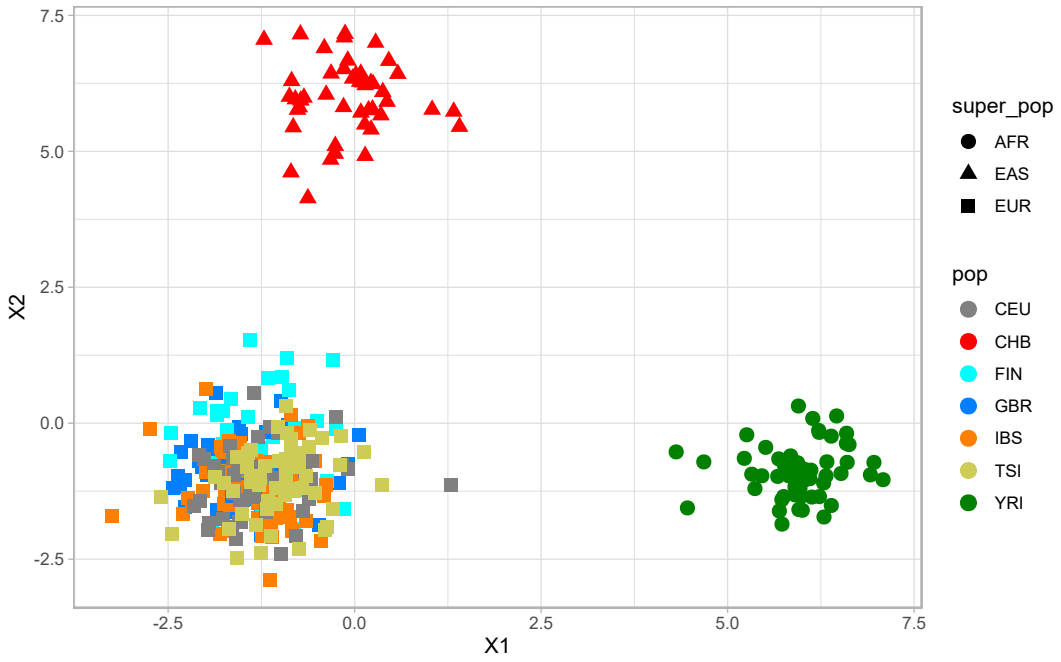

Figura 2.6: Representación de la varianza genética tras la aplicación del escalado multimensional sobre 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas.

2.3.3. Incrustación estocástica de vecinos indeterminados distribuidos en t (t-SNE).

El algoritmo de Incrustación Estocástica de Vecinos distribuidos en t o t-SNE (de *t-distributed Stochastic Neighbor Embedding*) es un método no lineal de reducción de dimensiones en la que los puntos vecinos en la nube de puntos son incrustados o embebidos utilizando una distribución t. A diferencia del método PCA, t-SNE preserva la distribución local de los datos permitiendo visualizar estructuras no lineales. El algoritmo fue introducido por Hinton and Roweis [24].

El algoritmo plantea que para cada individuo i, todo individuo j, tal que $i \neq j$, tiene una probabilidad p_{ij} de ser vecino. Esta probabilidad será:

$$p_{ij} = \frac{exp(-d_{ij}^2)}{\sum_{k \neq i} exp(-d_{ik}^2)} \tag{2.7}$$

siendo d_{ij}^2 la disimilitud entre los individuos y se calculará según la norma $\|\cdot\|$, definida en alta dimensión. La disimilitud entre dos puntos en alta dimensión x_i, x_j, se define como:

$$d_{ij}^2 = \frac{\|x_i - x_j\|^2}{2\sigma_i^2} \tag{2.8}$$

donde, σ_i se puede establecer manualmente.

Estos cálculos han de aplicarse en alta dimensión y su probabilidad homónima, q_i , en baja dimensión de tal forma que se minimice la función de pérdida que mide la divergencia de Kullback-Leibler [28].

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) ln(\frac{p(x)}{q(x)}) dx \tag{2.9}$$

Gracias a la minimización de valores de la ecuación 2.6 se puede representar en baja dimensión los individuos de un determinado estudio usando el algoritmo t-SNE (**Figura 2.7**).

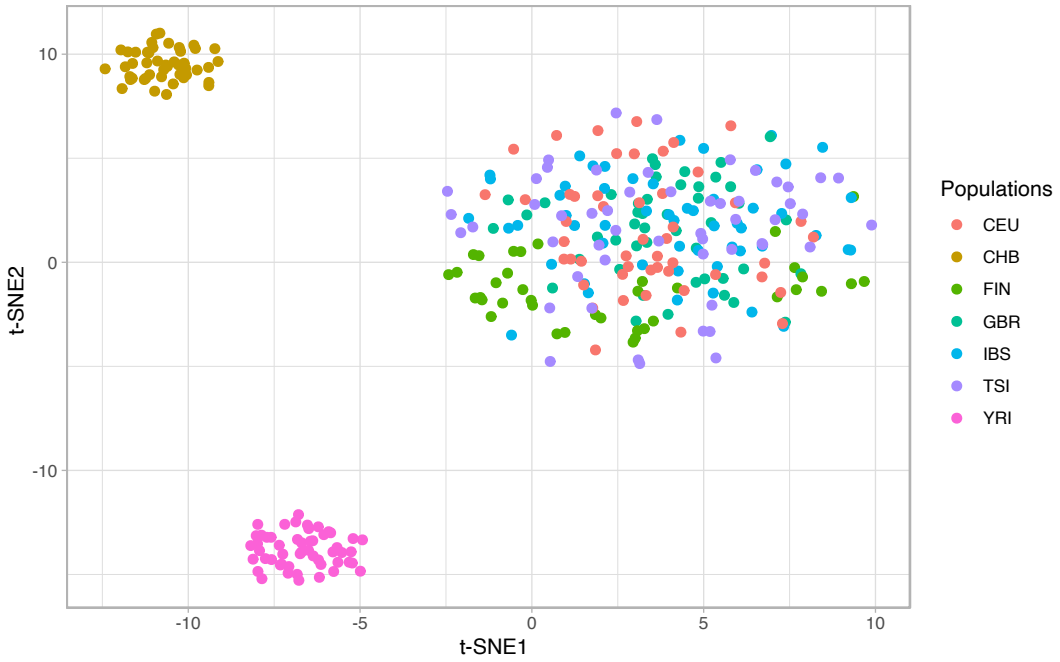

Figura 2.7: Representación de la probabilidad estocástica de vecinos en dos dimensiones tras la aplicación del método t-SNE a 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas.

2.3.4. Aproximación y Proyección Uniforme de Variedades (UMAP).

El algoritmo denominado Aproximación y Proyección Uniforme de Variedades o UMAP (de *Uniform Manifold Approximation and Projection*) es un método no lineal de reducción de dimensiones basado en técnicas de aprendizaje múltiple e ideas del análisis topológico de datos [25]. Debido a su eficiencia computacional, resulta adecuado para trabajar con datos masivos provenientes de tecnologías ómicas de alto rendimiento. UMAP se basa en construir un grafo de conectividad entre los puntos, de forma que establecen aristas entre los puntos cercanos. Luego, UMAP proyecta este grafo en un espacio de menor dimensión.

Se parte de una muestra de datos $X = X_1, ..., X_n$, que se suponen distribuidos uniformemente y se asume una métrica riemanniana, es decir, a que cada punto del espacio le asigna

una forma cuadrática definida positiva, que define su curvatura local, en su espacio tangente.

A partir de esta muestra de datos, se construye el grafo de conectividad definiendo la vecindad para cada punto en el espacio de alta dimensión usando la métrica definida riemanniana. Para definir el grafo correctamente, se calcula el peso de las aristas que se asignan entre dos vértices i y j en función de su distancia y la densidad local de puntos a su alrededor.

Se procede a proyectar el grafo en baja dimensión definiendo y optimizando una función de pérdida que mida la divergencia entre las distancias en el grafo original y las distancias entre los puntos proyectados en espacio de baja dimensión. Para ello, es común usar la divergencia de Kullback-Leibler definida en la **ecuación 2.9**.

En la **Figura 7** se muestra la aplicación del algoritmo UMAP con los siguientes parámetros: grafos de 15 vecinos, calculando dos componentes, usando la métrica euclídea y entrenando el modelo durante 200 épocas, sobre el mismo conjunto de datos obtenidos del Proyecto 1000 Genomas. Se observan tres agrupamientos similares a la reducción de dimensiones t-SNE. Sin embargo, estos agrupamientos presentan mayor compacidad.

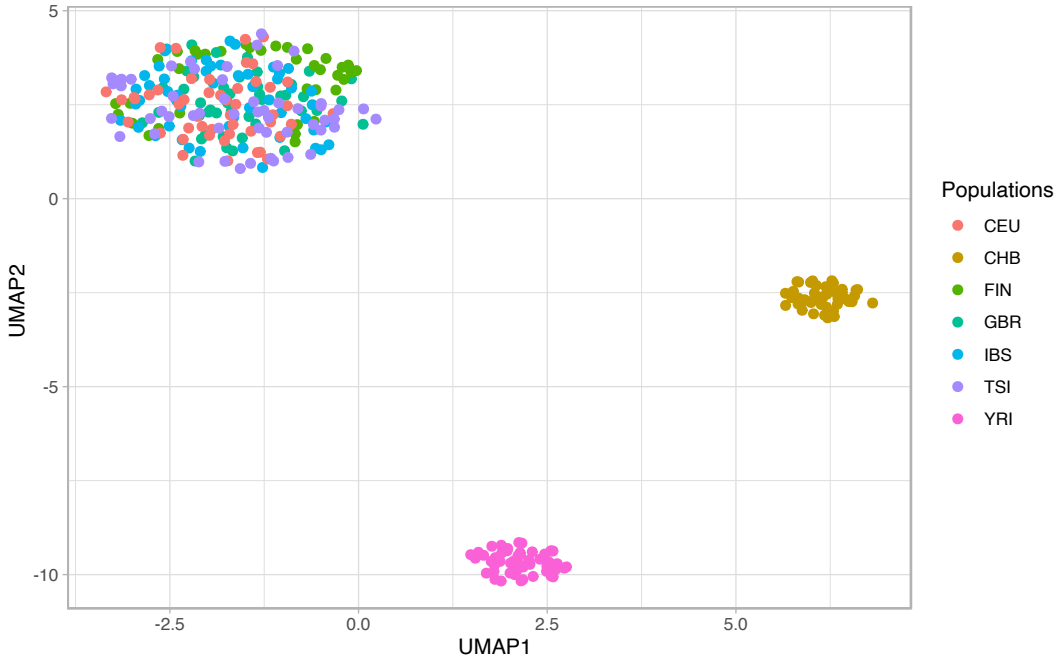

Figura 2.8: Representación en dos dimensiones usando el algoritmo UMAP para la reducción de dimensiones. Información genética de 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas.

2.4. Métodos de agrupamiento.

En el ámbito de la bioinformática, los métodos de agrupamiento se han convertido en herramientas esenciales para el análisis de grandes conjuntos de datos biológicos y la creación de

modelos predictivos (**Cuadro 2.3**). Existen dos tipos de métodos de agrupamiento: supervisado y no supervisado.

- **Supervisado.** Los algoritmos de agrupamiento supervisado permiten generar modelos de clasificación de datos biológicos. Para ello, en los datos ha de precisarse una variable dependiente u objetivo (comúnmente denominada *target*) que permitirá ajustar el modelo con el resto de variables.
- **No supervisado.** Los algoritmos de agrupamiento no supervisado permiten la clasificación de datos sin necesidad de definir una variable objetivo, sino por el descubrimiento de estructuras subyacentes.

Método	Etiquetado	Objetivo	Evaluación	Algortimos
Supervisado	Necesita una variable objetivo	Entrenamiento de modelos	Precisión del modelo	KNN, SVM
No supervisado	No necesita una variable objetivo	Descubrimiento de estructuras	Cohesión y separación	K-means, DBSCAN

Cuadro 2.3: Comparativa de las características de algunos métodos de agrupamiento supervisado y no supervisado. KNN , siglas en inglés de *K-Nearest Neighbors* algoritmo de aprendizaje supervisado ampliamente utilizado en el ámbito de la inteligencia artificial. SVM SVM, siglas en inglés de *Support Vector Machine* es un algoritmo de aprendizaje supervisado para tareas de clasificación.

Los métodos de agrupamiento no supervisado permiten descubrir patrones y relaciones no manifiestas *a priori* en datos genómicos, de expresión génica y proteómicos, entre otros, lo que resulta crucial para avanzar en la comprensión de procesos biológicos complejos y el desarrollo de nuevas estrategias en la detección de endofenotipos al realizar búsqueda de patrones entre los pacientes que son difíciles de encontrar de otra forma [9].

Existen distintos tipos de algoritmos para el agrupamiento no supervisado en función de qué características se observan cuando se aplica el algoritmo. Algunos de estos tipos son los siguientes:

- **Geométricos**. Este tipo de métodos divide el conjunto de datos en un número predefinido de grupos. El objetivo es encontrar una partición que minimice la varianza dentro de cada grupo y maximice la varianza entre los grupos.
- **Probabilísticos.** Este tipo de métodos usa modelos estadísticos para representar la distribución de los datos y descubrir las estructuras subyacentes en la nube de puntos.
- **Densidad relativa.** Estos métodos agrupan datos en función de la densidad de puntos en el espacio de características. La idea principal es identificar regiones densas en la nube de puntos que puedan interpretarse como agrupamientos.

En el **Cuadro 2.4** se presentan algunas de las ventajas e inconvenientes de los tipos algoritmos de agrupamiento no supervisado. El análisis en detalle de sus características permitirá una elección adecuada según el estudio que se pretenda desarrollar.

Algoritmos	Ventajas	Desventajas	Ejemplos
Geométricos	Simplicidad y eficiencia.	Requiere especificar número de agrupamientos.	K-means
	Fácil implementación.		K-medoides
	Útil para conjuntos de datos con formas de agrupamiento simples.	Sensibilidad a la inicialización de los centroides.	K-means++
Probabilísticos	Flexibilidad para modelar distribuciones complejas de datos.	Supuestos probabilísticos sobre la distribución.	Gausisan Mixture Models
	Útil para la detección de anomalías.	Requiere entrenamiento del modelo.	Mixture of Dirichlet proccess
Densidad relativa	Detección de agrupamientos de formas arbitrarias.	Requiere selección de parámetros adecuados.	DBSCAN
	Robustez.	Puede ser costoso su implentación para grandes conjuntos.	OPTICS

Cuadro 2.4: Análisis de las ventajas y desventajas de los tipos de algoritmos de agrupamiento no supervisado

En el caso del trabajo que se describe en esta memoria, se han usado principalmente tres algoritmos K-means, DBSCAN y HDBSCAN. A continuación, pasamos a descubrir en detalle cada uno de estos algoritmos.

2.4.1. K-means.

El algoritmo **K-means** es uno de los métodos de agrupamiento no supervisado más populares y utilizados. Este algoritmo pertenece a los agrupamientos de tipo geométrico (**Cuadro 2.4**), presentando tres características que lo convierten en un método muy útil: simplicidad, eficiencia y facilidad de implementación.

Este algoritmo tiene como objetivo la división de los datos en un **número predeterminado de grupos**. Para ello organiza los datos de forma que los puntos dentro de un mismo agrupamiento presenten características similares entre sí, al mismo tiempo que busca que los puntos pertenecientes a diferentes grupos sean lo más diferentes posible.

Formalmente, podemos definir la agrupación K-means como la búsqueda de la partición de datos en K grupos, $C := c_1, ..., c_k$ que minimiza la suma de cuadrados dentro del clúster (WCSS, *Within Clusters Sum of Squares*), utilizando:

$$WCSS_1 := \sum_{c_i \in C} \sum_{j=1 \cdots d} \sum_{x,y \in c_i} (x_{ij} - y_{ij})^2 \tag{2.10}$$

o su equivalente:

$$WCSS_2 := \sum_{c_i \in C} \sum_{j=1 \cdots d} 2|c_i| \sum_{x,j \in c_i} (x_{ij} - \mu_{ij})^2 \tag{2.11}$$

donde μ_{ij}, es la media de los puntos pertenecientes al agrupamiento i en dirección de la variable j, o dimensión matemática de j [29].

Para la actualización de los clusters, se calcula la nueva posición del centroide c_i como la media aritmética de todos los puntos de datos x_i asignados a ese agrupamiento. Este proceso se repetirá hasta llegar al número máximo de iteraciones establecidas, o hasta la estabilización de los centroides.

El algoritmo K-means utiliza un enfoque iterativo para minimizar la función objetivo y encontrar una partición óptima de los datos en K agrupamientos. La simplicidad del algoritmo y su base matemática sólida lo convierten en una herramienta de interés para el análisis de agrupamientos no supervisado en una amplia variedad de aplicaciones.

En la **Figura 2.9** se presenta el resultado de aplicar el algoritmo de agrupamiento no supervisado K-means para distintos valores de $K \in \{1, 2, 3, 4, 5, 6, 7\}$ a los datos genéticos de 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas (CEU, FIN, GBR, IBS, TSI, CHB y YRI) que se engloban, a su vez, en tres superpoblaciones (EUR, SAS y AFR).

A partir del valor $K = 4$ se observa que el algoritmo es capaz de proporcionarnos información sobre el agrupamiento de individuos en dimensiones superiores a las observadas en la figura. Esto es útil de cara a la identificación de características diferenciales dentro de un mismo grupo de muestras, observables por la reducción de dimensiones.

2.4.2. DBSCAN y HDBSCAN.

DBSCAN (*Density-Based Spatial Clustering of Application with Noise*), o Agrupamiento Espacial Basado en la Densidad de Aplicaciones con Ruido, y HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*), o Agrupamiento Espacial Basado en la

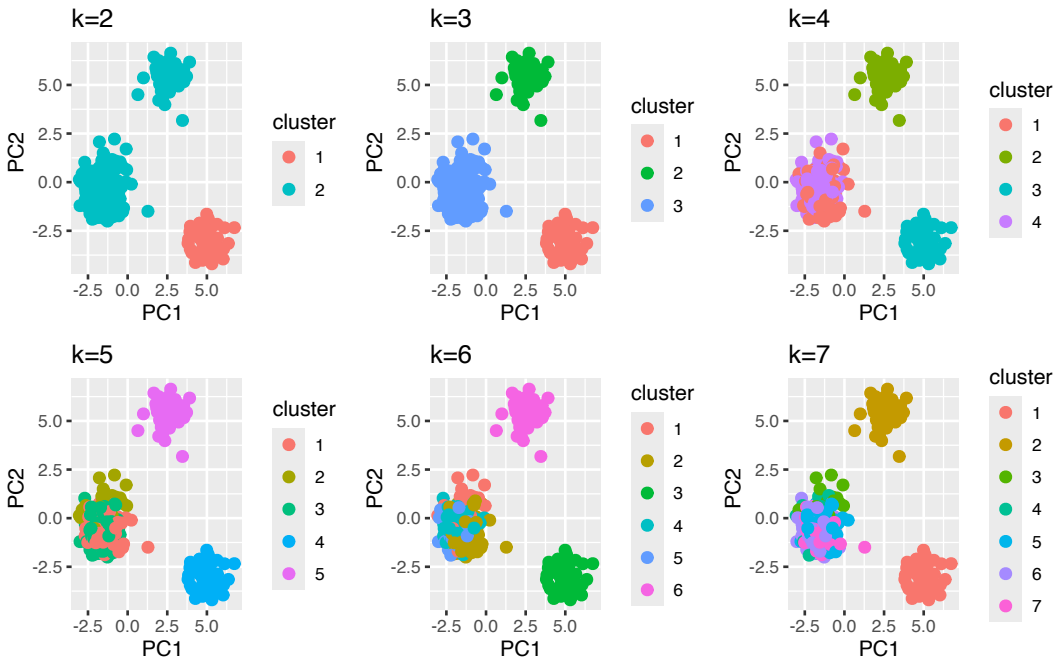

Figura 2.9: Aplicación del algoritmo K-means a los datos genéticos de 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas, tras haber aplicado previamente una reducción de dimensionalidad por PCA para proyectar los datos en dos dimensiones. Los agrupamientos descubiertos por el algoritmo K-means para cada valor de K se representan usando distintos colores.

Densidad Jerárquica de Aplicaciones con Ruido, son dos algoritmos de agrupamiento no supervisado basados en la densidad relativa de la nube de puntos. Estos algoritmos se utilizan para descubrir estructuras subyacentes en conjuntos de datos. Ambos métodos se dirigen a identificar regiones densas en la nube de puntos clasificándolos como un agrupamiento, mientras que los puntos de datos en áreas de menor densidad se consideran ruido o punto atípicos.

Ambos métodos presentan similitudes como algoritmos de agrupamiento basados en la densidad, pues permiten la detección de ruido y ninguno de ellos requiere que se deba especificar el número de agrupamientos previamente, como en el caso del algoritmo K-means.

Al mismo tiempo, ambos algoritmos emplean el parámetro *MinPts* para la determinación de la densidad mínima exigida para considerar una región como un agrupamiento. Para ello, se presupone un conjunto de datos en un espacio de variables y se define un radio de vecindad, ε. Para cada muestra, se crea una región o bola de vecindad de radio $r < \varepsilon$ que incluye todos los puntos dentro la misma.

La densidad de una muestra X se define como el número de muestras que se encuentran dentro de su región de vecindad. El parámetro *MinPts* ajusta la densidad mínima requerida para que una región se considere como agrupamiento, si y solo si la densidad de la región de vecindad de la muestra X es mayor o igual que *MinPts*, este se considerará parte de un agrupamiento.

En el **Cuadro 2.5**, se presenta una comparativa de las ventajas y desventajas de los algoritmos basados en densidad descritos en los apartados precedentes.

Algoritmos	Ventajas	Desventajas
DBSCAN	Simplicidad. Fácil implementación. Detección de agrupamientos con geometrías dispares	Sensibilidad al ruido. Computacionalmente costoso para su implementaicón en conjuntos de datos grandes.
HDBSCAN	Detección de agrupamientos jerárquicos. Menos sensibilidad al ruido. Más eficinete en conjuntos de datos grandes.	Mayor complejidad. Requiere ajustes adicionales.

Cuadro 2.5: Comparativa de algoritmos de agrupamiento basados en densidad: DBSCAN y HDBSCAN.

Analicemos cada uno de estos algoritmos por separado.

Algoritmo DBSCAN.

El algoritmo DBSCAN se puede resumir en los siguientes pasos [30]:

1. Encontrar los puntos en vecindad ε de cada punto, e identificar los puntos centrales con más vecinos que el parámetro *minPts*.
2. Encontrar los componentes conectados de los puntos centrales en el gráfico de vecinos, ignorando todos los puntos no centrales.
3. Asignar a cada punto no central a un agrupamiento cercano si el agrupamiento es un vecino de ε, de lo contrario se le identifica como ruido.

Para la optimización de cada iteración del algoritmo se introduce la función a optimizar. Para cada posible agrupamiento, $C = \{C_1, ..., C_l\}$ dentro del conjunto de todos los agrupamientos Γ, se minimiza el número de agrupamientos bajo la condición de que cada pareja de puntos en un agrupamiento es de *densidad alcanzable*, correspondiendo a las propiedades de maximización y conectividad de un agrupamiento:

$$\min_{C \subset \mathcal{C},\ d_{db}(p,q) \leq \varepsilon\ \forall p,q \in C_i\ \forall C_i \in C} |C| \tag{2.12}$$

donde $d_{db}(p, q)$ proporciona el mínimo tal que dos puntos p y q son de densidad alcanzable y $|C|$ es el cardinal del conjunto de agrupamientos, es decir, se busca el minimizar el número de

agrupamientos existente (**Figura 2.10**).

DBSCAN es una herramienta útil para el análisis de agrupamientos no supervisado, particularmente para la detección de clústeres de formas arbitrarias y en la identificación de ruido en los conjuntos de datos. Sin embargo, la elección del radio puede ser un desafío y, además, la sensibilidad al ruido puede ser un factor limitante.

Los resultados de la **Figura 2.10** se visualizan utilizando las dos primeras componentes del análisis de PCA (izquierda) y de UMAP (derecha), a partir de los genotipos de 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas ("0" indica "valor anómalo", señalado en color naranja).

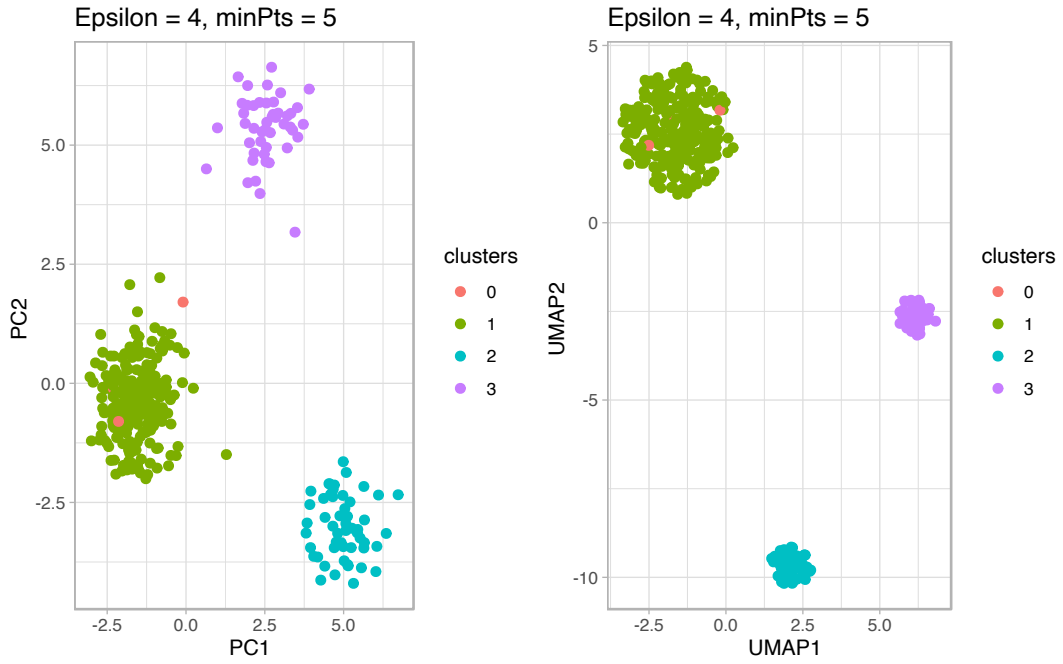

Figura 2.10: Agrupamientos identificados por DBSCAN ($\varepsilon = 4$; $minPts = 5$) aplicado sobre el resultado del análisis de componentes principales de los datos de demostración del Proyecto 1000 Genomas.

Se observa que la reducción de dimensiones UMAP muestra agrupamientos similares a los encontrados con DBSCAN. Sin embargo, considera ruido dentro del agrupamiento que puede ser debido a la elección de radio.

Una forma de abordar estas limitaciones con el radio y el ruido es utilizar el algoritmo HDBSCAN [25].

Algoritmo HDBSCAN.

HDBSCAN es una evolución del algoritmo DBSCAN. La principal característica de HDBSCAN es la introducción de la **densidad mutua inversa**, una medida de distancia que refleja la rareza de una muestra en función de su relación con su entorno. Esta medida permite a HDBSCAN identificar agrupamientos jerárquicos y anidados, detectando no solo agrupaciones sino también sus subagrupaciones y estructuras más finas dentro de los datos. De ahí que resulte de interés en la búsqueda e identificación de endofenotipos en los datos reales estudiados en este trabajo.

Para obtener la densidad mutua inversa, se calcula la **distancia de alcanzabilidad mutua**. La distancia de alcanzabilidad mutua de dos objetos x_p y x_q en el conjunto X se define como:

$$d_{am}(x_p, x_q) = max\{d_{nuc}(x_p), d_{nuc}(x_q), d(x_p, x_q)\} \tag{2.13}$$

Donde $d_{nuc}(x_i)$ es la distancia al núcleo definida como la distancia del objeto x_i a su *minPts*-vecino más cercano.

Los pasos principales del algoritmo HDBSCAN son los siguientes [31]:

1. Calcular la distancia al núcleo con respecto a *minPts* para todos los objetos de datos en el conjunto X.
2. Calcular el **grafo de alcanzabilidad mutua,** G, según la distancia de alcanzabilidad mutua.
3. Extender G añadiendo a cada vértice una arista que conecte con su mismo vértice con la distancia al núcleo del objeto como peso.
4. Extraer la jerarquía HDBSCAN como un dendograma:
   - *a*) Para la raíz del árbol, asignar a todos los objetos un mismo agrupamiento.
   - *b*) Eliminar iterativamente las aristas en orden decreciente de pesos.
     1) Antes de cada eliminación, se establece el valor de escala del dendograma del nivel jerárquico como peso de la arista a eliminar.
     2) Después de cada eliminación, se asignan como agrupamientos los componentes conectados que contienen los vértices finales de las aristas eliminadas para obtener el siguiente nivel jerárquico. Asignar un nuevo agrupamiento a un componente si tiene al menos una arista, si no, se le asigna la etiqueta nula ("ruido").

El grafo de alcanzabilidad mutua permite reconocer la densidad mutua inversa gracias a los pesos atribuidos a las aristas. De esta forma, el algoritmo es capaz de diferenciar los agrupamientos de manera jerárquica lo que permite identificar agrupamientos anidados de diferentes tamaños y densidades.

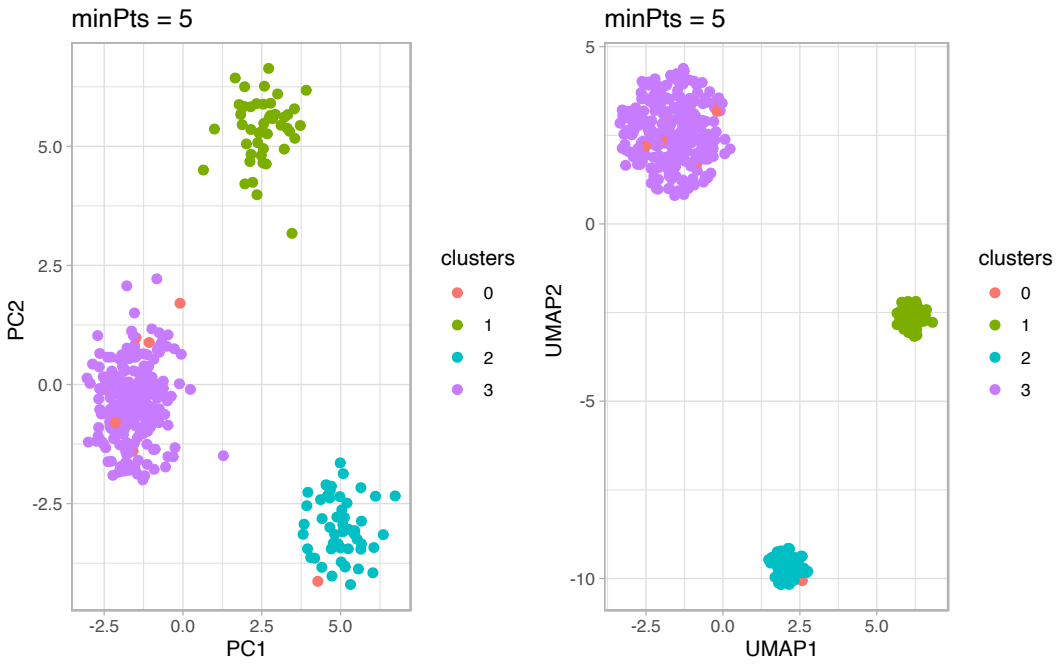

Figura 2.11: Agrupamientos identificados por HDBSCAN ($minPts = 5$) aplicado sobre el resultado del PCA.

Los resultados de la **Figura 2.11** se visualizan utilizando las dos primeras componentes del análisis de PCA (izquierda) y de UMAP (derecha), a partir de los genotipos de 715 individuos no relacionados de siete poblaciones del Proyecto 1000 Genomas ("0" indica "valor anómalo", señalados en color naranja).

En la gráfica de la representación de los datos del método de reducción de dimensiones UMAP se observan resultados similares a los encontrados en la aplicación del agrupamiento no supervisado HDBSCAN, detectando mayor ruido por la alcanzabilidad mutua.

Comparativa de los algoritmos de agrupamientos.

La elección entre ambos algoritmos depende de las características del conjunto de datos y el objetivo del análisis. Por ejemplo, si se desea medir las densidades de puntos que se encuentran definidos en bolas de un radio concreto se aplica DBSCAN. Si por el contrario, se desea medir densidades relativas en el espacio completo la aplicación de HDBSCAN obtiene un análisis completo.

En el desarrollo del trabajo se implementan ambos para facilitar la comparación de resultados obtenidos con los datos reales procedentes de los estudios en EPIDs y COVID-19 grave.

2.5. Entorno de trabajo.

Como parte del objetivo 1, y dado que el trabajo se desarrolla en colaboración con el Instituto Tecnológico y de Energías Renovables (ITER), el centro externo a la UOC con el que se ha desarrollado el trabajo, ha resultado conveniente crear un entorno de trabajo compartido

con el Tutor en la empresa. Para este menester se ha utilizado Google Drive y su capacidad para compartir y editar documentos en la nube.

La programación se ha realizado fundamentalmente en R, versión 4.3.3 [32], aunque se han realizado pruebas (no mostradas en esta memoria) en Python [33] utilizando las siguientes librerías: panda [34], para la gestión de los datos; keras [35], para la implementación de un marco de aprendizaje profundo o deep learning de alto nivel; y Matplotlib [36], para la visualización de resultados (desarrollo gráfico en 3 dimensiones).

Para la producción de los códigos en lenguaje R, se ha utilizado la interfaz gráfica de usuario RStudio [37].

Los principales paquetes de R utilizados se muestran en el **Cuadro 2.6** ordenados por su función más relevante en el contexto de este TFM.

Objetivo del paquete	Nombre del paquete	URL
Análisis exploratorio de los datos	tidyverse	[38]
	dummy	[39]
	MASS	[40]
	SNPRelate	[41]
Reducción de dimensiones	umap	[42]
	Rtsne	[43–45]
Agrupamiento (Clustering)	dbscan	[46, 47]
	ConsensusClusterPlus	[48]
	stats	[32]
Interfaz de usuario	shiny	[8]
	shinythemes	[49]
Visualización	ggplot2	[50]
	plotly	[51]
	gridExtra	[52]

Cuadro 2.6: Paquetes de R utilizados en el TFM.

El control de versiones se ha realizado utilizando un repositorio privado de GitHub disponible en https://github.com/. Se ha dejado en modo público en el momento de depósito del presente trabajo.

2.6. Productos obtenidos.

Para finalizar el apartado materiales y métodos, se comentarán brevemente los productos obtenidos en el desarrollo del TFM.

Los productos tangibles se muestran en el repositorio de GitHub del TFM: https://github.com/. Para el uso correcto del repositorio se ofrece un documento MarkDown **README.md** en el que se resume el objetivo y la metodología usada para el desarrollo del repositorio, así como a quién se dirige principalmente este repositorio y los archivos contenidos en él.

El repositorio consta de tres carpetas con objetivos distintos. A saber:

- La carpeta RMD está destinada al desarrollo de un informe RMarkDown realizado sobre el estudio de datos de proteómica de alto rendimiento en EPID y está pensado para el análisis y agrupamiento de datos por métodos no supervisados. Esta herramienta permitirá la replicación de estudios similares gracias al dinamismo de este documento.
- La carpeta Script contiene un total de seis archivos que describen la metodología seguida en el archivo RMarkDown. De este se pueden extraer métodos de R para el análisis de variables y de conjuntos de datos, para la preparación de datos de proteómica de alto rendimiento (Olink) y, en definitiva, para la aplicación de métodos de agrupamiento no supervisado sobre un conjunto de datos. Además todo el código se ha comentado el inglés para aumentar el alcance del mismo.
- Por último, en la carpeta UI se presenta un script de R que permite el acceso a la interfaz de usuario programada en shiny. Esta interfaz de usuario facilita la aplicación de métodos de agrupamiento no supervisado en datos ya tratados, permitiendo a otros usuarios realizar un estudio similar al presentado en este TFM.

Capítulo 3

Resultados y discusión.

El presente capítulo presenta los resultados obtenidos en el marco del TFM, cuyos objetivos generales fueron descritos en el **Apartado 1.2**. del **Capítulo 1**.

Como se indica en el **apartado 2.2**, se ha programado y aplicado un flujo de trabajo bioinformático completo. Los resultados de este TFM se describen gracias a la aplicación del flujo de trabajo bioinformático sobre los datos procedentes de los dos estudios referidos anteriormente.

Los resultados referentes al objetivo 1, la búsqueda bibliográfica, se encuentran detallados en el **Capítulo 2**, en el cual se describen los métodos estudiados para su aplicación en los diferentes datos que vienen proporcionados por la búsqueda bibliográfica. Sin embargo, se ha decidido incluir a continuación una descripción del estado actual en la investigación, *Estado del arte*, como resultado del objetivo 1.

3.1. Estado del arte

Las enfermedades respiratorias crónicas son una de las causas principales de muerte y discapacidad en el mundo [53]. Por esta razón, la Organización Mundial de la Salud instauró en mayo del año 2000 la *Global Alliance against Chronic Respiratory Disease*, GARD. En el informe [54] de la última reunión global del GARD, en diciembre de 2023, se señaló la necesidad de desarrollar más investigación en los métodos de diagnóstico de enfermedades pulmonares obstructivas, como pueden ser las EPIDs.

En el proceso de diagnóstico de enfermedades, los biomarcadores tienen un papel determinante, pues facilitan la identificación rápida de procesos desencadenantes de enfermedades o se asocian con ellos. Según el National Cancer Institute o NCI [55], un biomarcador es una: *"molécula biológica que se encuentra en la sangre, en otros fluidos corporales o en los tejidos y que es signo de un proceso normal o anormal, o de una afección o enfermedad."*

Este hecho ha provocado que numerosas investigaciones se centren en el desarrollo tecnológico que permita un mayor acceso a la obtención de datos de proteómica de alto rendimiento, conjuntos de información a gran escala que describen las proteínas presentes en una muestra

biológica de los individuos estudiados [12].

En concreto, las EPIDs presentan grandes dificultades en su diagnóstico y en su tratamiento [56]. Recientemente se ha realizado una investigación proteómica en sangre sobre la FPI [57]. Esto ha permitido identificar y validar un total de 140 proteínas biomarcadoras que se asocian con la supervivencia de los pacientes. Estos resultados dan cuenta del enorme potencial de la proteómica en las EPIDs.

La búsqueda de biomarcadores de enfermedades respiratorias también tiene relevancia en el estudio de endofenotipos de COVID-19 grave. Un reciente estudio sugiere que las vías inflamatorias específicas relacionadas con el daño tisular están implicadas en la aparición de endofenotipos de COVID-19 grave [58].

Para el desarrollo de la investigación de los datos de proteómica se han implementado métodos de análisis de datos con *Machine Learning* [12]. Ambas disciplinas, proteómica y *machine learning*, están a la vanguardia de la investigación biológica.

El *machine learning* o aprendizaje automatizado es la implementación de algoritmos que permitan al ordenador realizar una tarea concreta. En el contexto de la bioestadística y el análisis multivariante, el aprendizaje automático implementa algoritmos estadísticos sofisticados para identificar patrones complejos.

Los algoritmos desarrollados pueden ser supervisados o no supervisados. Los algoritmos supervisados son aquellos que tienen una fase de entrenamiento para encontrar patrones que permitan la diferenciación de dos o más grupos, por ejemplo para distinguir en una muestra aleatoria de mujeres, aquellas que padecen anorexia nervosa [59]. Los algoritmos no supervisados son aquellos que aprenden a partir de un conjunto de datos, sin etiquetar, descubriendo patrones, agrupaciones o estructuras ocultas en la información sin necesidad de intervención humana.

La profundidad y el poder de los algoritmos de *machine learning* vienen acompañados de una exigente necesidad de abstracción matemática. Para comprender este campo es necesario adentrarse en conceptos necesarios para el análisis multivariante [60]. Entre estos conceptos se encuentran los espacios vectoriales en los que se definen los pacientes, matrices que compactan la información del estudio o representan aplicaciones matemáticas, las distribuciones de probabilidad multivariante y la optimización de soluciones a problemas complejos.

En definitiva, en las publicaciones más recientes de proteómica se han aplicado algoritmos de *machine learning* para la distinción de EPIDs asociadas a la enfermedad del tejido conectivo (CTD) respecto de la FPI [14]. También estas dos disciplinas se han combinado para la detección de biomarcadores de EPIDs [61].

3.2. Agrupamiento no supervisado en pacientes de EPID.

Se comienza por realizar un análisis exploratorio univariante de las muestras de pacientes con EPID. Las variables que acompañan a los datos de proteómica de alto rendimiento son cuatro:

- *Site:* Lugar de obtención de la muestra (discreta).
- *Sex:* Sexo del paciente (discreta).
- *Ancestry*: Ascendencia del paciente (discreta).
- *Age*: Edad en el momento de obtención de la muestra (continua).

Además, a parte de estas cuatro variables se estudia como fenotipo el diagnóstico del paciente.

3.2.1. Análisis univariante.

Se realiza un primer análisis univariante de las variables y fenotipo de las muestras. En las **Figura A.1** se muestra un diagrama de barras de las variables discretas estudiadas. Se incluyen otras figuras de interés en el **Anexo A**

Respecto a la variable fenotípica, se reconoce una amplia mayoría de casos de diagnóstico 1, por lo que se decide usar este diagnóstico como contraste. De esta manera tendremos los individuos que han sido diagnosticados con la enfermedad 1 y los individuos que han sido diagnosticados con otra enfermedad considerados como individuos de control.

Además de las variables discretas, se dispone de una variable de distribución continua, la edad, que se ha representado usando dos gráficas que se muestran en la **Figura 3.1**.

Se observa que la mayor parte de los pacientes diagnosticados con EPIDs se encuentran entre los 60 y 80 años, tal y como era de esperar a la vista del impacto tienen las EPIDs en los pacientes de estas edades.

Para continuar con el análisis univariante de las variables, se procede a analizar la distribución de los datos de las proteínas. En la **Figura 3.2** se muestra la densidad de cuatro proteínas escogidas al azar entre el total de 368 proteínas estudiadas.

Dado el carácter asimétrico y apuntado de las distribuciones de densidad y con el fin de extraer el máximo de información posible, se han estimado los momentos estadísticos de orden superior: el sesgo y la kurtosis (ver **Anexo B**) de las distribuciones. Estos estadísticos permiten intuir la densidad de las proteínas sin necesidad de visualizarlas

Como resultado del análisis de la kurtosis, se ha observado que un total de 329 proteínas, es decir, un 89 %de las proteínas, presenta una densidad leptocúrtica. Las 39 proteínas restantes,

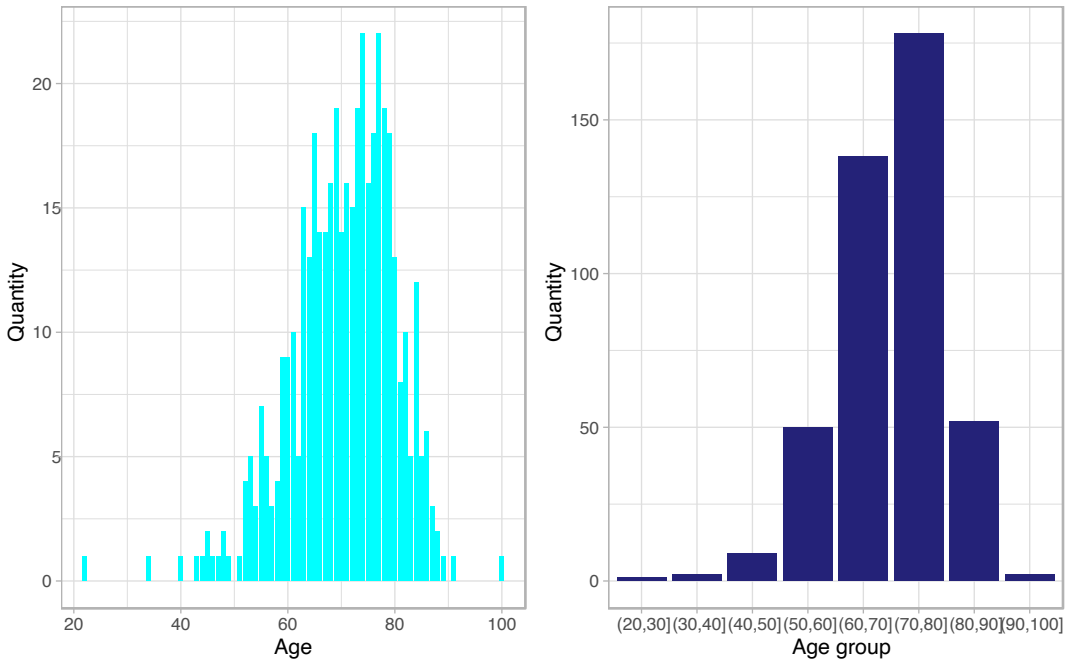

Figura 3.1: Diagramas de la variable edad. Datos absolutos (izquierda) y datos estratificados por décadas de edad (derecha).

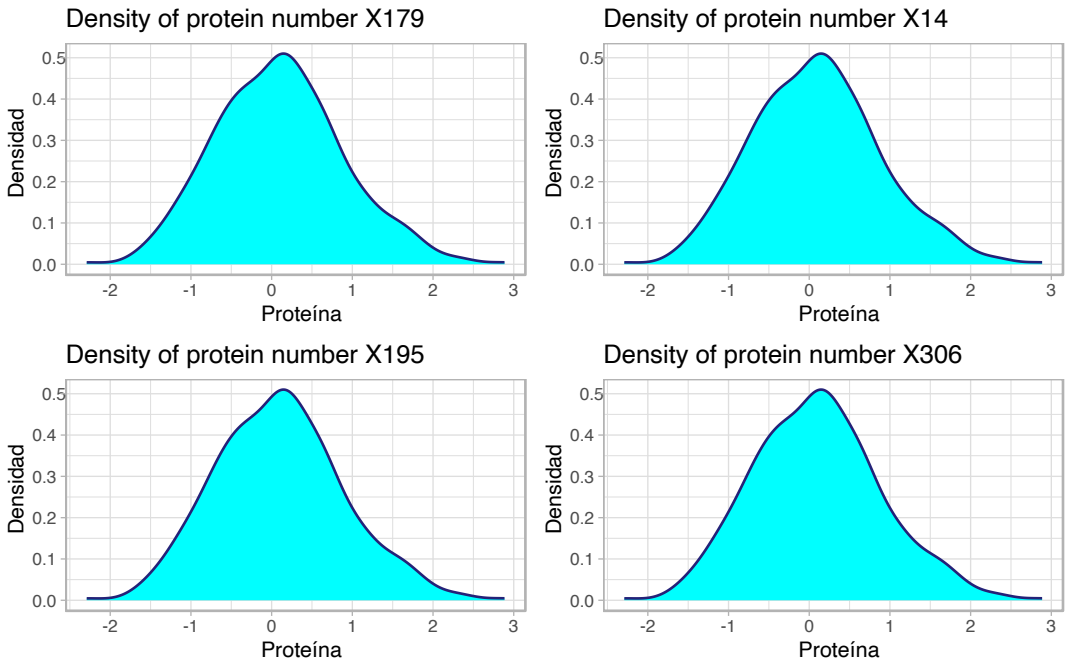

Figura 3.2: Distribución de los valores de expresión proteómica de las proteínas 179, 14, 195 y 306, escogidas aleatoriamente entre el total de proteínas del estudio.

el 11 % de las proteínas, presentan una distribución platicúrtica.

Respecto al resultado del sesgo, el 100 % de las proteínas presentan asimetría con respecto a los valores mayores que su media.

En el estudio univariante de las proteínas también se realiza el diagrama de cajas, mostrado en la **Figura** ° del **Anexo A**, se usan las mismas proteínas que se escogieron aleatoriamente para las distribuciones de densidad.

Los diagramas de cajas permiten la detección de valores atípicos de manera univariante, debido al objetivo del estudio se desea conservar el máximo de muestras, por lo que se decide no eliminar los valores atípicos de forma univariante.

3.2.2. Análisis bivariante.

Para el análisis bivariante de las proteínas, se realiza el estudio de la correlación de las proteínas usando el método de Pearson. En las figuras del **Anexo A** se muestra un histograma de los valores que presenta la matriz de correlaciones.

Los valores de correlación obtenidos permiten discernir entre las parejas de proteínas que presentan valores de expresión relacionados. En la **Figura 3.3** se muestra la regresión lineal de 9 parejas de proteínas aleatorias que muestran una alta correlación ($r > 0{,}95$), probablemente porque se trata de proteínas relacionadas biológicamente. Se observa la relación de dependencia entre los valores observados de las proteínas.

Por otro lado, si el valor absoluto de la correlación entre dos proteínas es cercano a 0 se constata que no existe esa dependencia. En la **Figura A.3**, se muestra la regresión lineal parejas de proteínas con bajo índice de correlación.
Las proteínas de la **Figura A.3** presentan una correlación superior a $0,01$ que es el umbral definido como baja correlación. En las figuras se observa que no existe relación lineal, pero además de la relación lineal las proteínas podrían presentar relaciones de otro tipo (polinómicas, cuadráticas, logarítmicas, etc.). Sin embargo, la nube de puntos de la comparación de proteínas muestra que no existe ninguna de esas relaciones.

3.2.3. Análisis multivariante.

El análisis multivariante se ha enfocado en la detección de valores atípicos. Para ello se ha calculado la distancia de Mahalanobis (**Ver Anexo C**) con estimadores robustos del vector media y de la matriz de covarianzas de los perfiles proteicos de los individuos.

En la **Figura 3.5** se muestra la distribución de las distancias de Mahalanobis Cuanto mayor es el valor en el eje Y, el individuo se encuentra a mayor distancia del vector media y, por tanto, se considera como valor atípico.
Se ha seleccionado un umbral que permita identificar las muestras atípicas del estudio usando la desviación típica,σ , de las distancias calculadas. Se ha marcado como valores atípicos aquellos individuos que muestren un perfil proteómico arroje una distancia de Mahalanobis

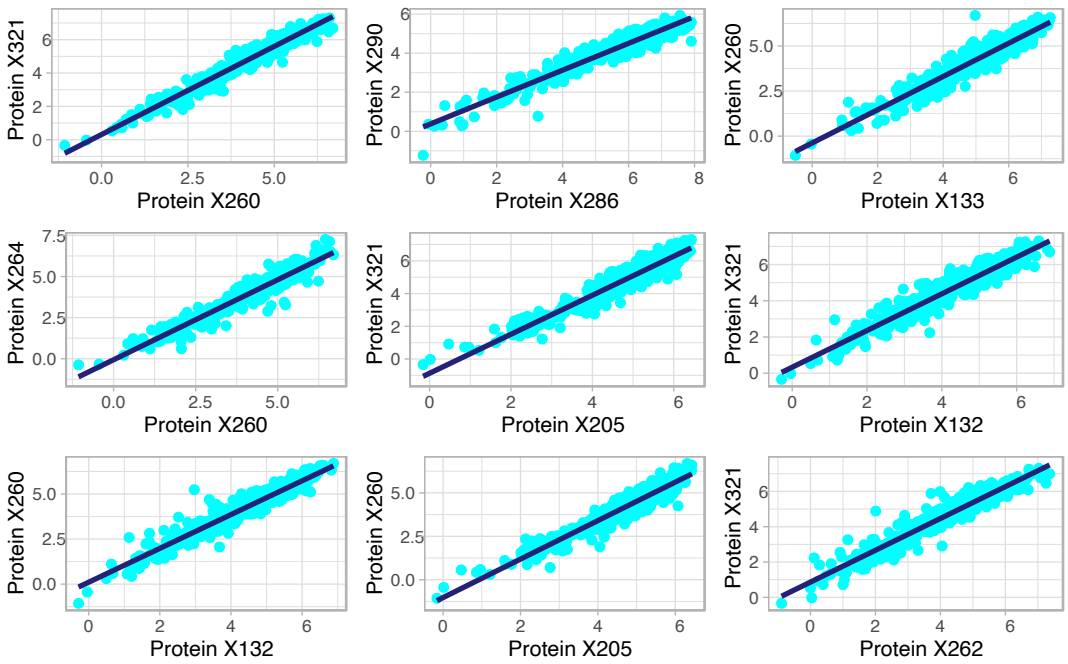

Figura 3.3: Regresión lineal de 9 parejas de proteínas escogidas aleatoriamente que presentan correlación de Pearson superior a un umbral definido. Las proteínas de la figura presentan una correlación $r > 0,95$.

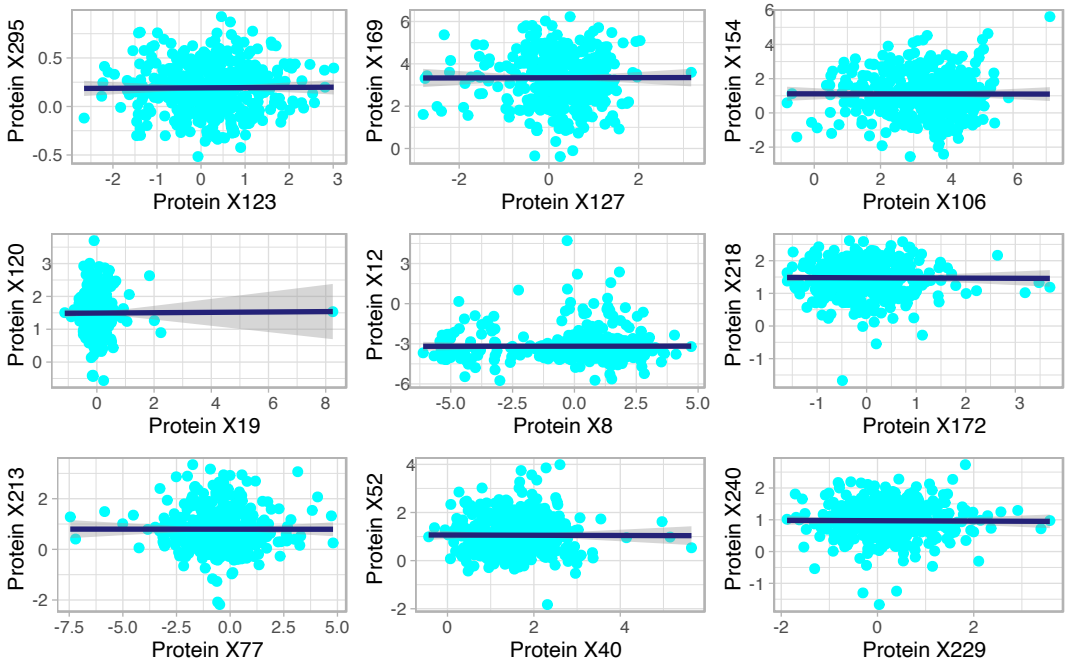

Figura 3.4: Regresión lineal de 9 parejas de proteínas escogidas aleatoriamente que presentan una correlación de Pearson inferior a un umbral definido ($r < 0,01$).

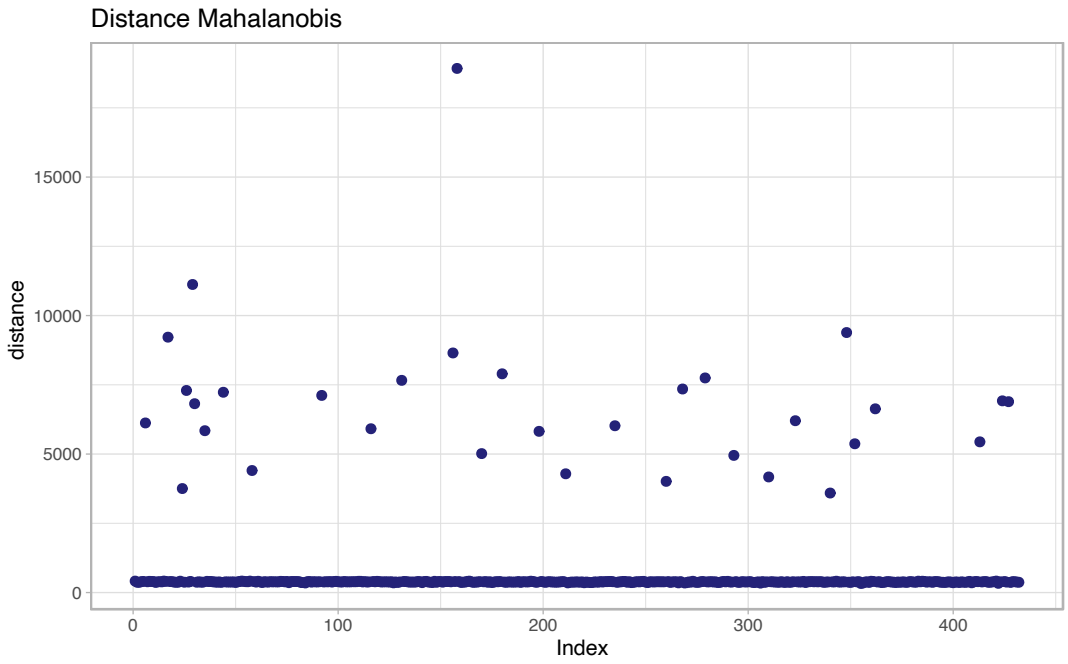

Figura 3.5: Nube de puntos de individuos tras la aplicación de la distancia Mahalanobis.

superior a 3σ.

El **Cuadro 3.1** recoge los datos de las muestras consideradas como atípicas. Estas muestras no se eliminarán del estudio. Sin embargo, se marcarán durante la aplicación de los métodos de reducción de dimensiones y los agrupamientos realizados.

3.2.4. Reducción de dimensiones.

En este apartado se muestran los resultados al aplicar los tres métodos de reducción de dimensiones detallados en el **apartado 2.3** de este documento. El PCA se ha utilizado para la definición de los individuos en el espacio de proteínas. Esto ha permitido la reducción de covarianza al definir a los individuos a partir de una base que no presenta colinealidad. El resto de métodos de reducción de dimensiones se ha aplicado en los datos para su representación en dos dimensiones.

En el **apartado 2.2** se han explicado las diferentes matrices usadas para los métodos de reducción de dimensiones. En la **Figura 3.6**se observa la reducción de dimensiones por UMAP de dos matrices de datos. En esta reducción se ha usado las opciones por defecto del método (métrica euclídea, número de vecinos 15 y épocas 200).

Ambas gráficas se han coloreado según el diagnóstico del individuo, comparando las muestras de diagnóstico 1 con las muestras de otros diagnósticos.

Una vez aplicado el método UMAP para la reducción de dimensiones se procede a su comparación con la reducción de dimensiones por la aplicación del método t-SNE sobre las mismas

Identificador	Diagnóstico	Centro	Sexo	Edad	Ascendencia
S1894	1	1	1	74	W
S0006	2	2	2	62	W
S0213	1	2	2	71	W
S0233	2	2		60	W
S0246	3	2	1	65	W
S0273	1	1	1	52	H
S0286	1	1	1	70	A
S0292	1	1	1	80	W
S0301	2	1	1	75	B
S0323	1	1	1	68	W
S0347	1	1	1	68	W
S0354	3	1	1	52	B
S0355	4	1	1	63	W
S0366	4	1	2	80	W
S0368	4	1	1	68	A
S0374	1	1	1	68	W
S0400	1	1	1	77	W
S0408	4	1	1	75	W
S0058	1	4	1	71	W
S0065	1	4	2	67	W
S0086	1	4	2	67	W
S0096	1	4	1	76	W
S0104	1	4	1	84	W
S0184	1	4	1	84	W

Cuadro 3.1: Información fenotípica de los individuos que superan el umbral definido como valor atípico ($> 3\sigma$).

matrices de datos. La aplicación de este método se observa en la **Figura 3.7**.

Ambas gráficas de la **Figura 3.7** se han coloreado según su diagnóstico, comparando las muestras de diagnóstico 1 con las muestras de otros diagnósticos.

Al comparar las **Figura 3.6** y la **Figura 3.7** resulta que la aplicación de ambos métodos de reducción de dimensiones no aporta información adicional en el caso de los datos de proteómica de alto rendimiento, puesto que no se aprecian diferencias apreciables a simple vista en la representación de los datos.

Por esta razón, durante el desarrollo de los agrupamientos no supervisados se decide usar el método UMAP como método junto con MDS para la representación de los datos por su similitud con los métodos no supervisados que se han aplicado en el **apartado 3.2.5**.

Como última reducción de dimensiones, se ha aplicado el método del escalamiento multidimensional con la distancia Minkowski de potencia 1 en la **Figura 3.8**.

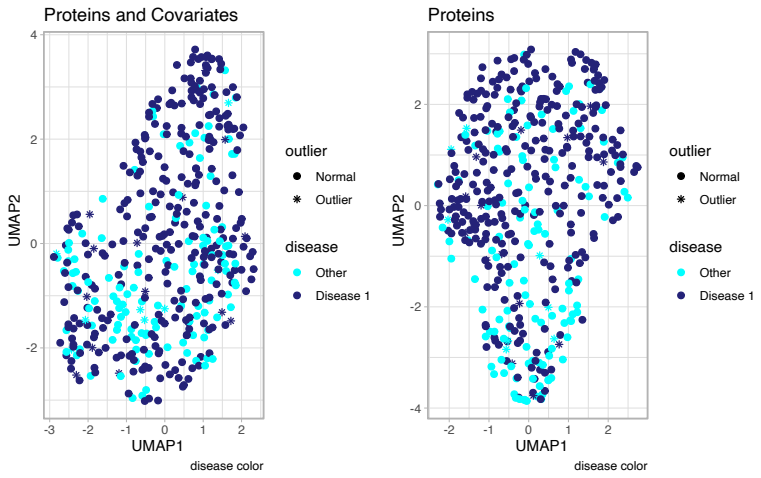

Figura 3.6: Reducción de dimensiones por el método UMAP de la matriz de datos de expresión proteínas y covariables (izquierda) y la matriz de expresión de proteínas (derecha).

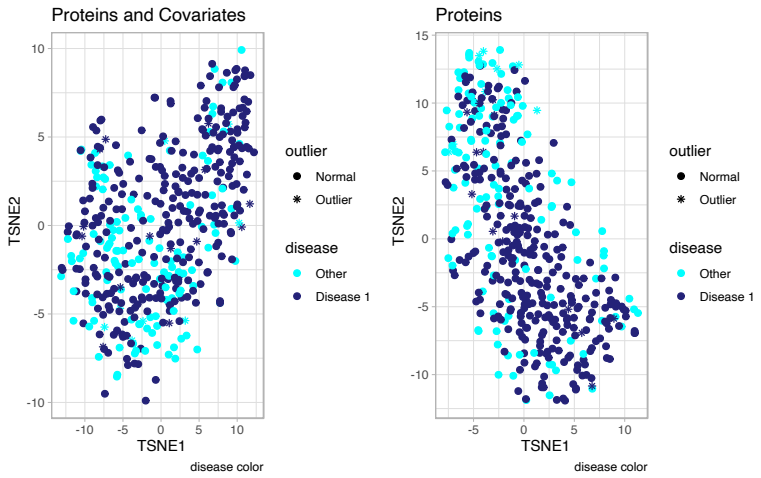

Figura 3.7: Reducción de dimensiones por el método t-SNE en los datos de pacientes con EPID contenidos en la matriz de proteínas y covariable (izquierda) y solo en la matriz de proteínas (derecha).

Las dos gráficas de la Figura 24 se han coloreado según la variable definida como fenotipo, *diagnóstico*, comparando las muestras de *diagnóstico 1* con las muestras marcadas con otros diagnósticos.

La decisión de usar como métrica la distancia Minkowski de potencia 1, o distancia Manhattan, es debida a su equivalencia con la métrica euclídea, la métrica usada para los agrupamientos no supervisados. Sin embargo, las distancias definidas por la métrica Minkowski de potencia 1 presentan valores mayores.

A continuación, se procede con la aplicación de agrupamientos no supervisados.

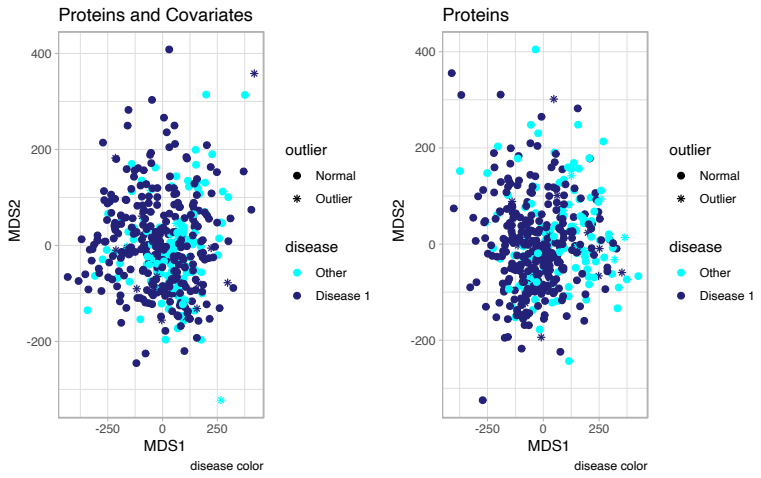

Figura 3.8: Reducción de dimensiones por el método MDS de los datos de pacientes con EPID sobre la matriz de distancias de proteínas y covariable (izquierda) y solo la matriz de distancias de proteínas (derecha).

3.2.5. Agrupamiento no supervisado.

Una vez se ha realizado la reducción de dimensiones y se conoce mejor la distribución de los datos en un espacio que es comprensible por su baja dimensión, se procede a la aplicación de los métodos de agrupamiento no supervisado. Se empieza por la aplicación del método K-means con $k = 2$, en las matrices *data* y *proteins*.

La **Figura 3.9** muestra los agrupamientos encontrados tras la aplicación de este método sobre las matrices de datos estudiadas usando el método de reducción de dimensiones UMAP.

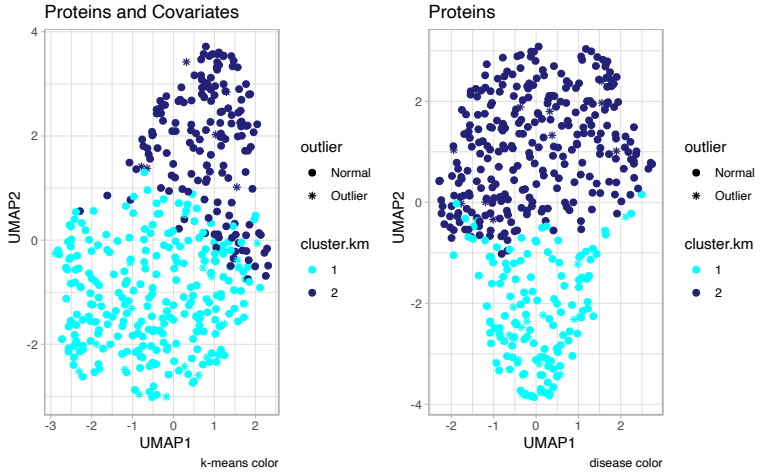

Figura 3.9: Aplicación de K-means, de $k = 2$. Visualización con UMAP. Se muestran los resultados de la aplicación de K-means sobre los datos de proteínas y covariables (izquierda) y sobre los datos de proteínas exclusivamente (derecha).

Ambas gráficas usan el método de reducción de dimensiones no lineal UMAP. Se observa que existe un gran cambio en los agrupamientos dependiendo de la matriz estudiada.

Además, en ambas gráficas se comprueba que la combinación del uso de un algoritmo de agrupamiento no supervisado lineal y un método de reducción de dimensiones no lineal dificulta la visualización de la bola definida por el método K-means. Para solventar estas dificultades a continuación, se procede a mostrar el mismo agrupamiento usando MDS como método de reducción de dimensiones (**Figura 3.10**).

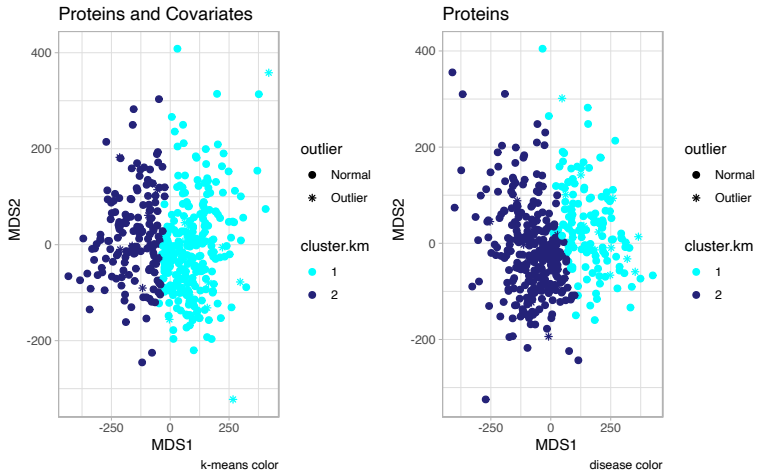

Figura 3.10: Aplicación de K-means, de$k = 2$. Visualización con MDS. Datos de proteínas y covariables (izquierda) y datos de proteínas (derecha).

Ambas gráficas usan el método de reducción de dimensiones MDS. Se constata que el método MDS muestra hiperplanos mejor definidos en la representación de los datos en baja dimensión.

En la **Figura 3.11** se muestra el mismo agrupamiento de datos usando la matriz de datos de proteínas y covariables, y se compara el método de reducción de dimensiones.

En la comparación de ambos métodos de reducción de dimensiones, se concluye que al reducir la dimensión de los datos por el método MDS se distingue un hiperplano que divide los datos.

Una vez se comparan resultados del agrupamiento no supervisado geométrico (K-means) sobre ambos métodos de reducción de dimensiones, se procede a la comparación del agrupamiento con el fenotipo que se desea estudiar. La **Figura 3.12** y la **Figura A.4** del **Anexo A** muestran la comparación de los datos según el fenotipo estudiado frente al agrupamiento realizado usando el método K-means.

De manera análoga, en la **Figura 3.13** y la **Figura A.5** del **Anexo A** se muestran la comparación de los datos según el fenotipo estudiado frente al agrupamiento realizado usando el método K-means.

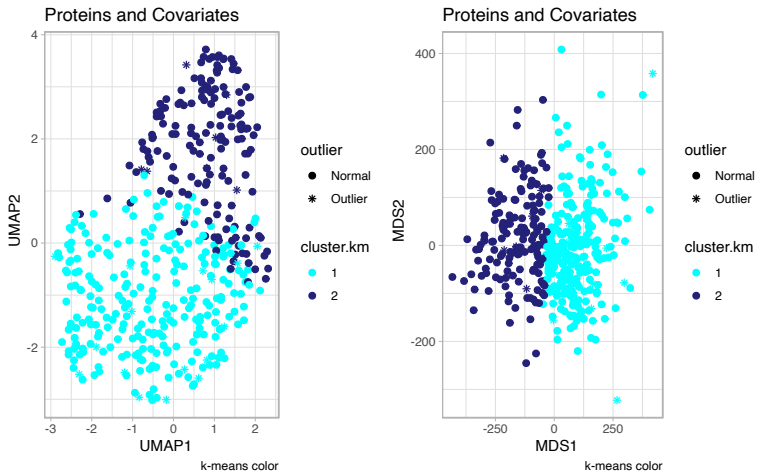

Figura 3.11: Aplicación de K-means, de $k = 2$. Comparación de métodos de reducción de dimensiones. Se utilizan los datos de proteínas y covariables en ambas gráficas. En este caso, se ha utilizado el método de reducción de dimensiones UMAP (izquierda) y el método de reducción de dimensiones MDS (derecha).

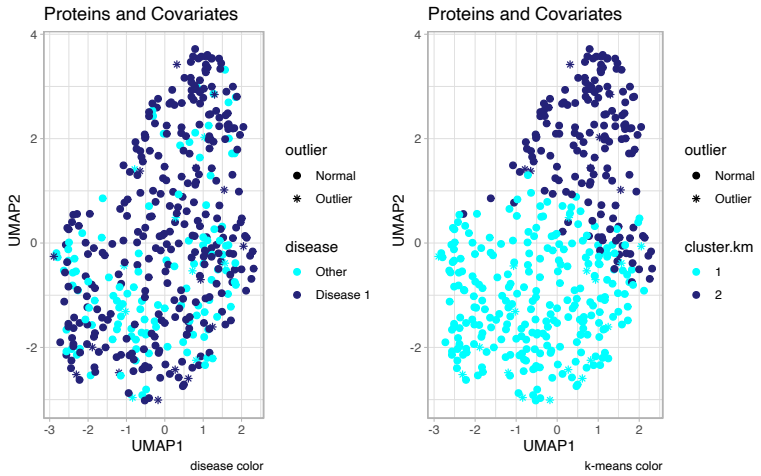

Figura 3.12: Representación de los datos de proteínas y covariables con el método de reducción de dimensiones de UMAP coloreado según el diagnóstico (izquierda). Representación de los datos de proteínas y covariables con el método de reducción de dimensiones de UMAP coloreado según el método de agrupamiento no supervisado a la (derecha).

Se observan diferencias notables entre los datos de proteómica coloreados según el fenotipo del estudio (enfermo / no enfermo) y el agrupamiento realizado por el método K-means. Se prosigue el estudio con la aplicación de otros métodos de agrupamiento no supervisado basados en las densidades.

Se procede a la aplicación de DBSCAN como método de agrupamiento no supervisado. En la **Figura 3.14** se muestran los resultados de la aplicación del método de agrupamiento no supervisado DBSCAN en los datos de proteínas y covariables y en los datos de proteínas.

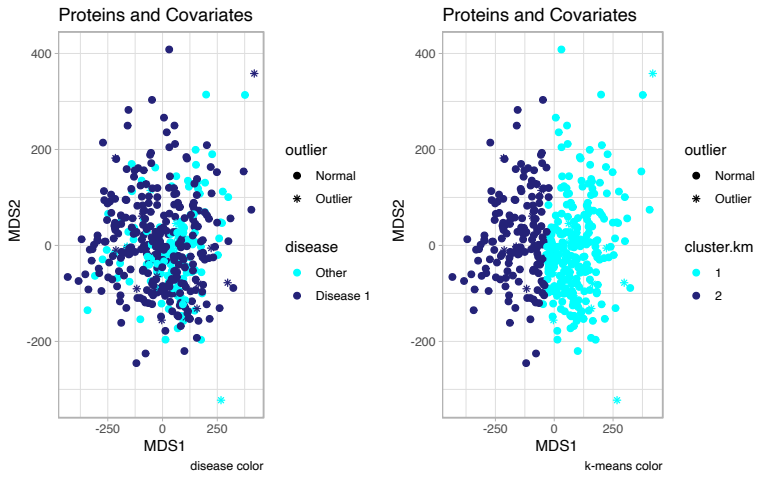

Figura 3.13: Representación de los datos de proteínas y covariables con el método de reducción de dimensiones de MDS coloreado según el diagnóstico (izquierda). Representación de los datos de proteínas y covariables con el método de reducción de dimensiones de MDS coloreado según el método de agrupamiento no supervisado (derecha).

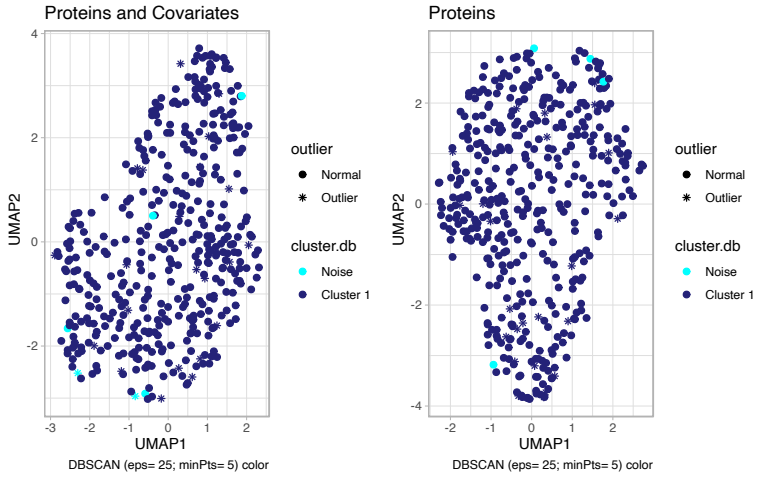

Figura 3.14: Aplicación de DBSCAN, de $\varepsilon = 25$ y $minPts = 5$, sobre los datos de proteínas y covariables (izquierda) y sobre los datos de proteínas (derecha). Ambas gráficas usan el método de reducción de dimensiones no lineal UMAP.

La distribución de los datos en el espacio de alta dimensión no permite a DBSCAN, con los parámetros seleccionados, la detección de agrupamientos. En la **Figura 3.15** se muestra la aplicación del método de agrupamiento no supervisado DBSCAN con el método de reducción de dimensión MDS.

Al combinar estos dos métodos, se constata que no existen muestras de densidades de la distribución en el espacio aleatorio de al menos 5 puntos en un radio de 25 unidades y, por lo

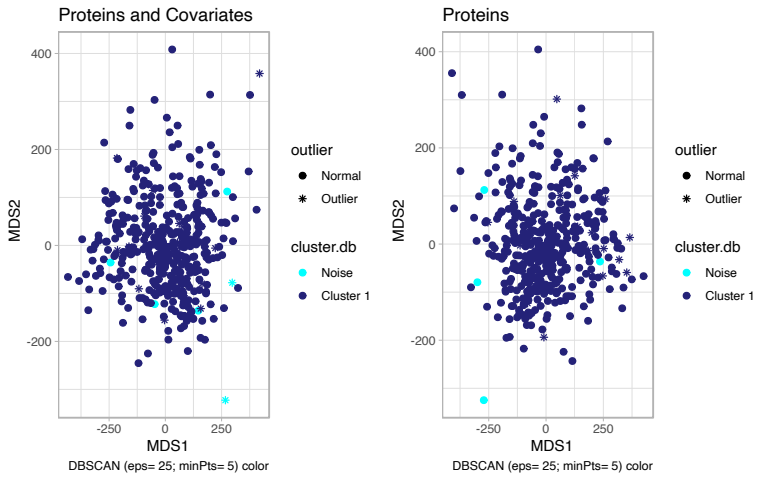

Figura 3.15: Aplicación de DBSCAN, de $\varepsilon = 25$ y $minPts = 5$, sobre los datos de proteínas y covariables (izquierda) y sobre los datos de proteínas (derecha). Ambas gráficas usan el método de reducción de dimensiones MDS.

tanto, el algoritmo DBSCAN solamente es capaz de detectar un único agrupamiento y muestras que considera ruido. Por lo tanto, se procede a la aplicación del método HDBSCAN.

En la **Figura 3.16** se muestra la aplicación del método de agrupamiento HDBSCAN en los conjuntos de datos de proteínas y covariables, por un lado, y de proteínas, por otro.

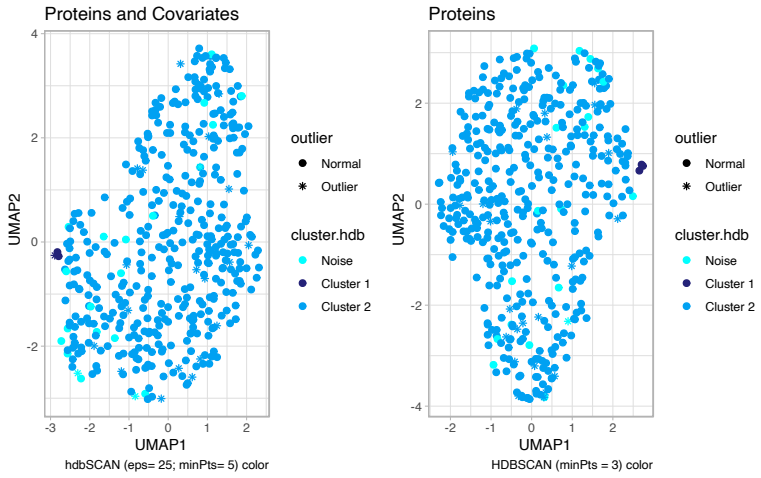

Figura 3.16: plicación de HDBSCAN, de $minPts = 3$, sobre los datos de proteínas y covariables (izquierda) y sobre los datos de proteínas (derecha). Ambas gráficas usan el método de reducción de dimensiones no lineal UMAP.

En la **Figura 3.16** se observa que el método HDBSCAN ha sido capaz de ajustar un radio de capaz de encontrar dos agrupamientos diferentes. Los **Cuadros 3.2 y 3.3** muestran los datos fenotípicos de los individuos detectados como Cluster 1 en el borde de la nube de puntos

proyectada en baja dimensión.

Identificador	Diagnóstico	Sitio	Sexo	Edad	Ascendencia
S0231	4	1	1	79	W
S0265	4	1	1	61	W
S0286	1	1	1	70	A

Cuadro 3.2: Datos fenotípicos de las muestras identificadas en el *Cluster 1* en la aplicación del método de agrupamiento no supervisado HDBSCAN en la matriz de datos de proteínas y covariables.

Identificador	Diagnóstico	Sitio	Sexo	Edad	Ascendencia
S0051	1	4	1	79	W
S0081	1	4	1	80	W
S0102	1	4	1	75	W

Cuadro 3.3: Datos fenotípicos de las muestras identificadas en el Cluster 1 en la aplicación del método de agrupamiento no supervisado HDBSCAN en la matriz de datos de proteínas.

En este caso, se han identificado dos agrupamientos de los datos distintos según la matriz de datos estudiada. En el **Capítulo 4** se procede a la extracción de conclusiones de los resultados obtenidos en este apartado.

3.3. Agrupamiento no supervisado en pacientes de COVID-19.

En este caso se ha realizado un análisis exploratorio univariante de los datos de pacientes con COVID-19 grave tratados con corticoides. Las covariables de este estudio son las siguientes:

- *Sex:* Sexo del paciente (discreta).
- *Age:* Edad en el momento de obtención de la muestra (continua).
- *Center:* Lugar de obtención de la muestra (discreta).

Aparte de estas covariables, se dispone como fenotipo la mortalidad a 90 días del paciente.

3.3.1. Análisis univariante.

En las **Figura A.7** del **Anexo A** se muestra la distribución de las variables discretas empleando diagramas de barras.

Ninguna de las tres variables (sexo, edad y mortalidad) se distribuye uniformemente. A continuación, se ha realizado el análisis univariante de la edad de los pacientes (**Figura 3.17**).

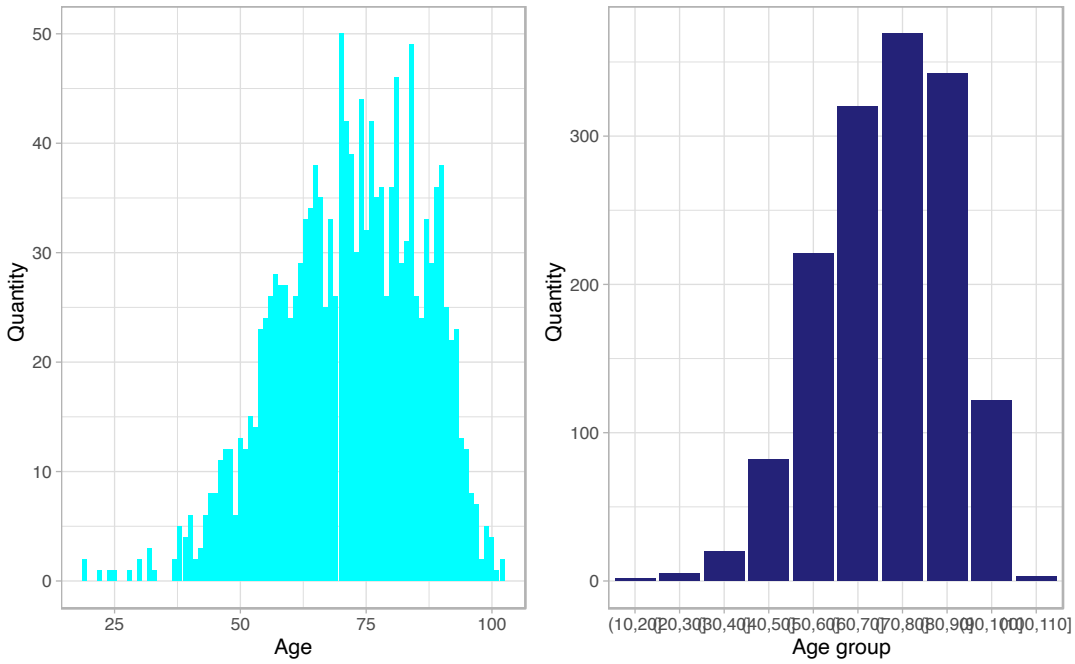

Figura 3.17: Histograma (izquierda) de la edad y estratificación de la edad mediante diagrama de barras (derecha) de los pacientes de COVID-19 grave tratados con corticoides.

La distribución de la edad de los pacientes con COVID-19 grave refleja una mayor frecuencia entre los 60 y 90 años, observándose la moda entre los 70 y 80 años. Como es sabido, los pacientes con edades en este intervalo se encuentran en la conocida como *población de riesgo*.

La *población de riesgo* está formada por los individuos que tienen mayor probabilidad de sufrir daños graves al contraer la enfermedad y, por tanto, es previsible que su mortalidad pueda ser mayor. No obstante, se advierte de que en este conjunto de datos no se dispone de datos de comorbilidades (que se espera sean de relevancia a medida que la edad de los pacientes crece).

El estudio univariante de los datos continúa con el análisis de las componentes principales que representan la estructura genética de los individuos del estudio (se han obtenido a partir de un conjunto de aproximadamente 100,000 variantes independientes). En la **Figura 3.18** se presentan las gráficas de densidad de cada una de las componentes principales estudiadas.

Las componentes principales se distribuyen normalmente. El cálculo del sesgo y la kurtosis señala que todas las componentes principales son leptocúrticas, es decir, todas ellas poseen picos más prominentes que la distribución normal, y todas son asimétricas con colas a la derecha.

Como se dispone de los autovalores o *eigenvalues* obtenidos en el PCA, se ha confeccionado el denominado *scree-plot* (**Figura 3.19**). Este diagrama representa la variación explicada por cada componente principal.

Del scree-plot se desprende que la primera componente principal explica una mayor variación

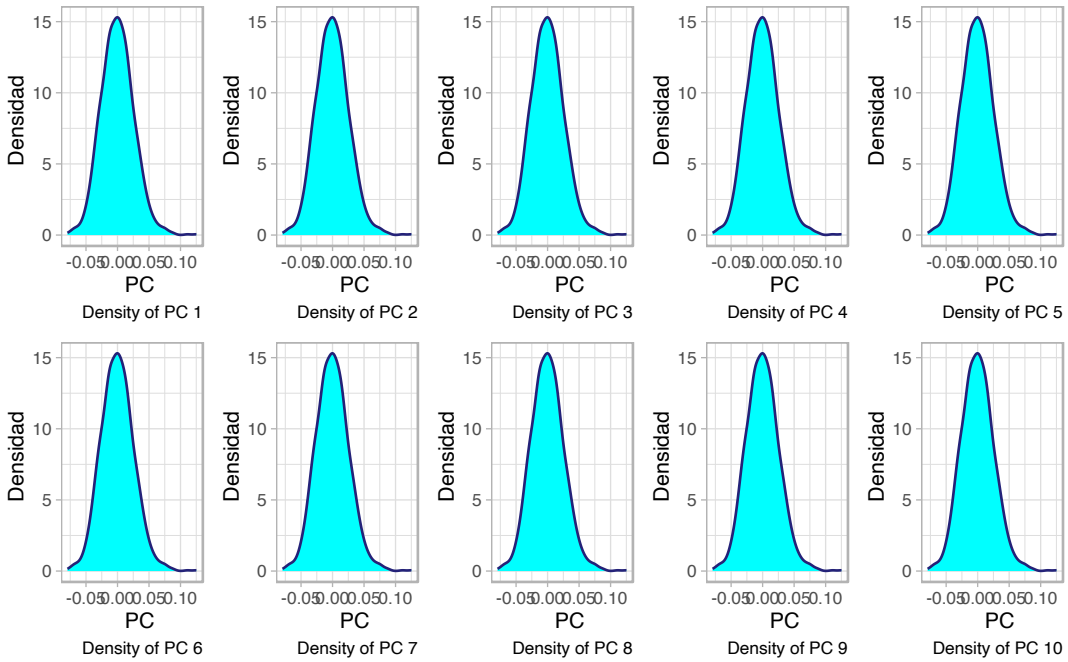

Figura 3.18: Gráfico de densidad de los componentes principales de los genomas de los pacientes de COVID-19 grave tratado con corticoides.

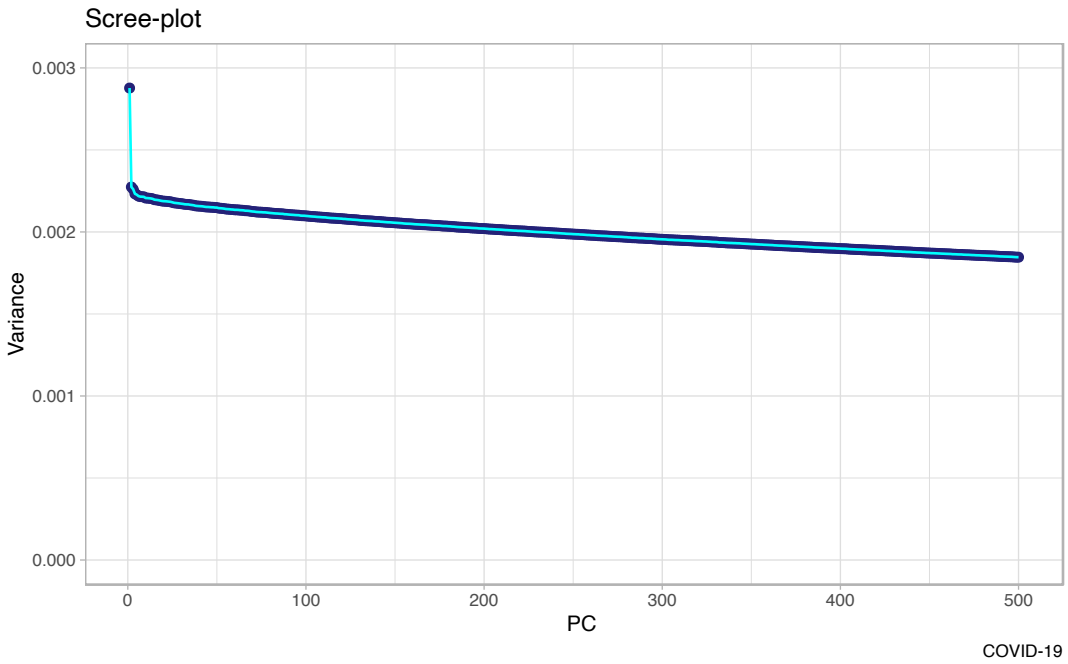

Figura 3.19: Scree-plot de los autovalores asociados a las componentes principales de los pacientes de COVID-19 grave tratado con corticoides.

en los datos genéticos de los individuos. A continuación, encontramos la segunda componente principal, y así sucesivamente. La diferencia entre la variabilidad explicada por la primera

componente principal y la última es de aproximadamente 0,07 %.

3.3.2. Análisis bivariante.

El análisis exploratorio de los datos se complementa con un análisis bivariante. Para ello se ha realizado la regresión lineal entre la primera componente principal y la edad, diferenciada según el sexo del individuo (**Figura 3.20**).

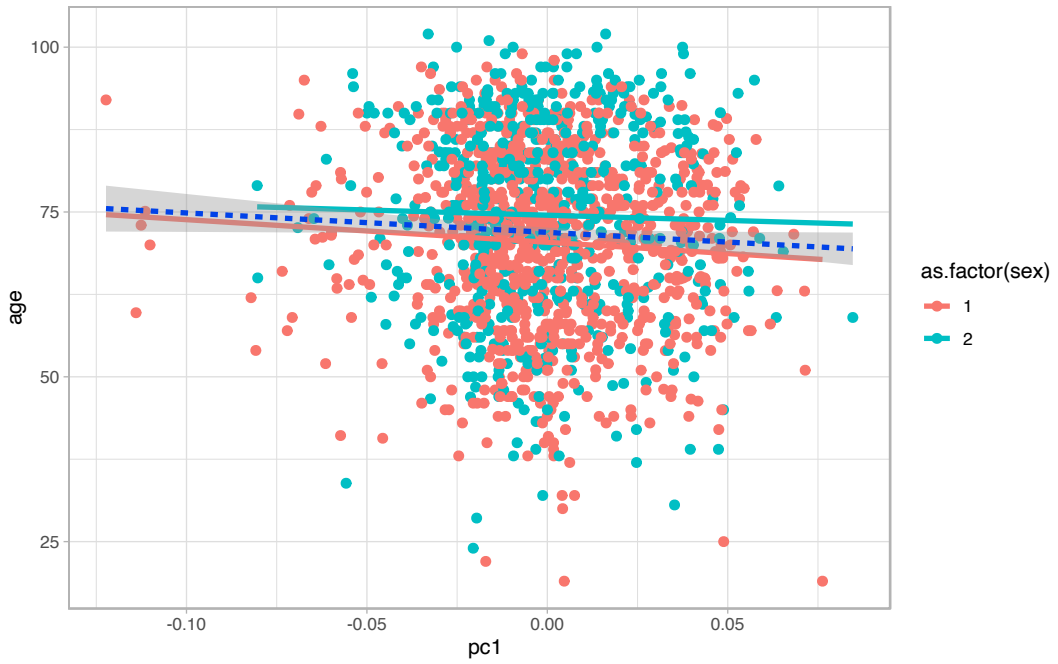

Figura 3.20: Regresión lineal de la primera componente principal y la edad según el sexo.

De este diagrama se desprende que no existe relación lineal, ni relación polinómica dada la nube de puntos, entre la primera componente principal, la edad y el sexo.

3.3.3. Análisis multivariante.

El análisis multivariante se ha centrado en la detección de valores atípicos por medio de estimadores robustos del vector media y la matriz de covarianzas. Se ha empleado la distancia de Mahalanobis para la determinación de estos valores.

En la **Figura 3.21** se presentan las distancias de los individuos al vector media de las diez primeras componentes principales del estudio.

Como se observa en la **Figura 3.21**, algunos pacientes quedan representados a gran distancia de la media. Esta observación nos permite definir como valores atípicos aquellos pacientes que superen un umbral definido como 3 veces la desviación estándar de las distancias a la media.

El **Cuadro 3.4** muestra los datos de los valores atípicos encontrados en el estudio.

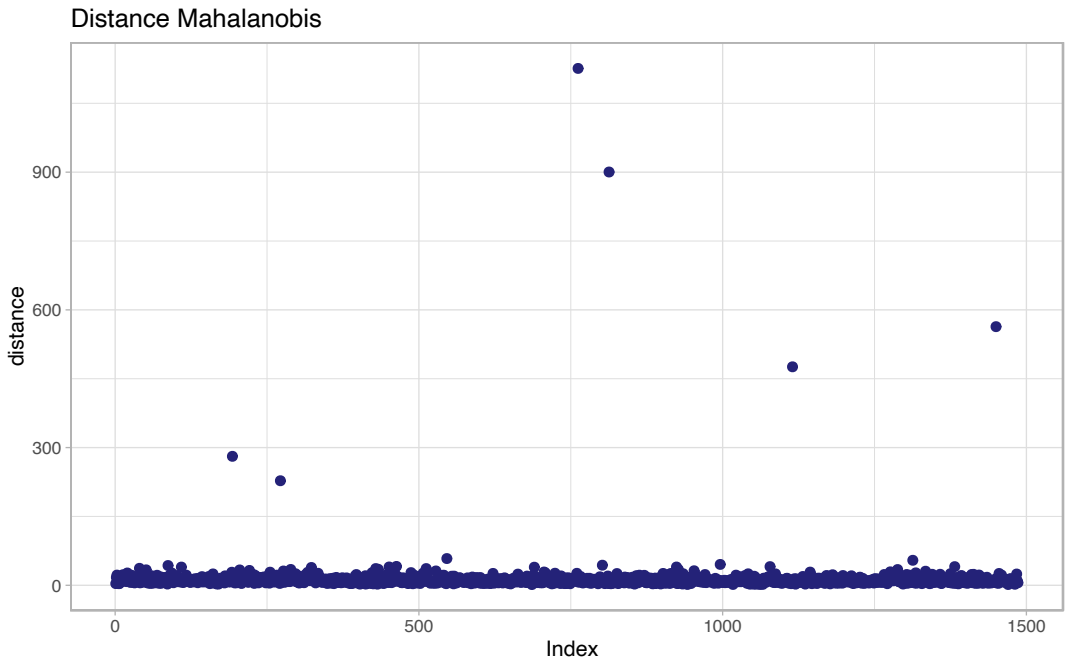

Figura 3.21: Distancia de Mahalanobis de los pacientes con COVID-19 grave tratados con corticoides.

Paciente	Sexo	Edad	Centro	Mortalidad
196	2	79	3	1
276	2	93	4	2
771	2	92	8	2
822	1	57	8	1
1125	1	64	8	1
1464	1	63	8	1

Cuadro 3.4: Valores atípicos de los pacientes con COVID-19 grave tratados con corticoides.

Al igual que en el estudio realizado con los datos de proteómica de alto rendimiento, no se eliminarán los valores atípicos de este segundo estudio.

3.3.4. Reducción de dimensiones.

A continuación, se han aplicado los tres métodos de reducción de dimensiones a las matrices indicadas en el **apartado 2.2**. Aplicamos el método MDS para la reducción de dimensiones. En este caso, se ha aplicado usando la métrica de Minkowski de potencia 1. Los resultados se observan en la **Figura 3.22**. En todos los casos que se exponen a continuación se visualizan los resultados de la reducción de dimensiones coloreando los individuos según la mortalidad a 90 días.

La reducción de dimensiones de la **Figura 3.22** sobre la matriz de distancias de los componentes principales muestra la cercanía genética de los individuos. En la reducción realizada

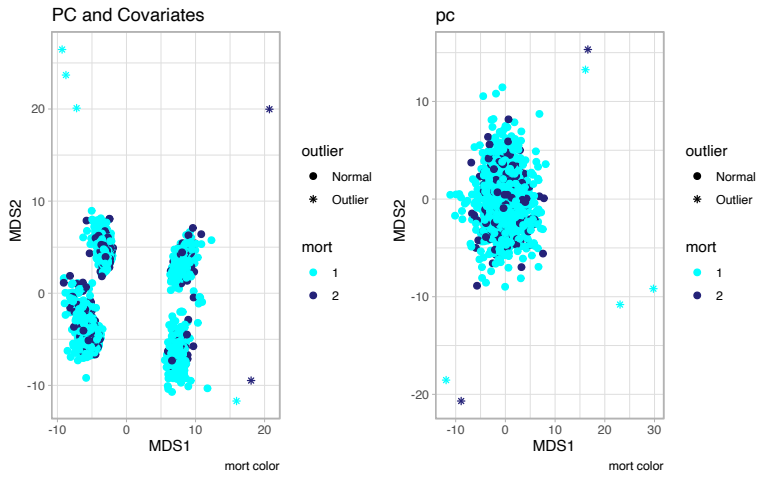

Figura 3.22: Reducción de dimensiones aplicando el método MDS de los datos de pacientes con COVID-19 grave tratado con corticoides utilizando la matriz de distancia de los componentes principales y covariables (izquierda) y solo la matriz de distancia de componentes principales (derecha).

sobre la matriz de distancias de componentes principales y covariables se muestran agrupamientos de individuos en alta dimensión.

A continuación, se ha utilizado el algoritmo de UMAP en R, con el resultado mostrado en la **Figura 3.23**.

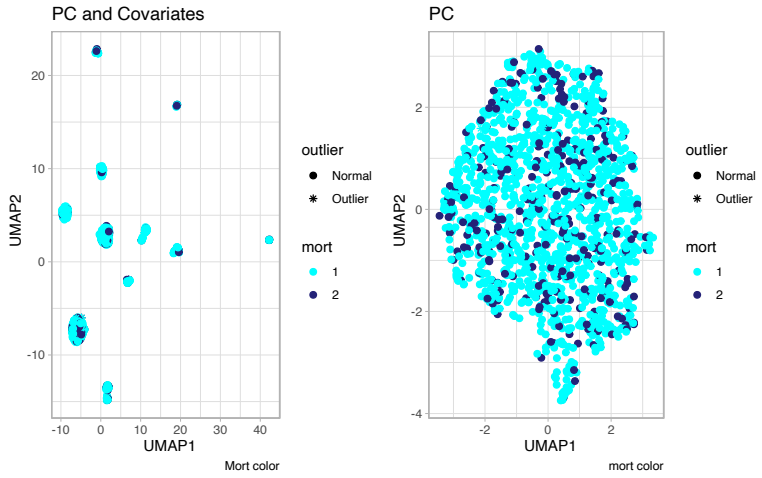

Figura 3.23: Reducción de dimensiones aplicando el método UMAP a pacientes con COVID-19 grave tratados con corticoides utilizando la matriz de componentes principales y covariable (izquierda) y solo la matriz de componentes principales (derecha).

Al usar el método UMAP se comprueba la relación mencionada anteriormente observada con el método MDS. Además, la aplicación del método UMAP sobre la matriz de componentes

principales y covariables muestra más agrupamientos de individuos que el método MDS.

En la **Figura 3.24** se presenta la reducción de dimensiones que resulta de aplicar el método t-SNE.

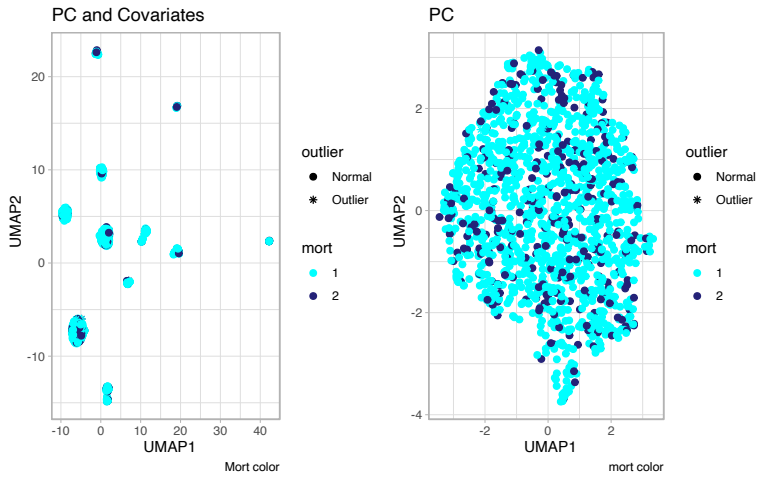

Figura 3.24: Reducción de dimensiones aplicando el método t-SNE de los datos de pacientes con COVID-19 grave tratado con corticoides utilizando la matriz de componentes principales y covariables (izquierda) y solo la matriz de componentes principales (derecha).

En este caso, se observa una gran semejanza con la forma de agrupar los datos de UMAP. Sin embargo esta representación presenta bolas de individuos de mayor amplitud facilitando la visualización del color para distinguir los agrupamientos.

Se procede a la aplicación de métodos de agrupamiento no supervisado.

3.3.5. Agrupamiento no supervisado.

Para la realización del agrupamiento no supervisado se ha comenzado por aplicar el método K-means (**Figura 3.25**) sobre las componentes principales visualizando el resultado utilizando el método de reducción de dimensiones de UMAP.

La comparación de agrupamientos encontrados al aplicar K-means en ambas matrices de datos muestra la sensibilidad que tiene el algoritmo a definir un agrupamiento dependiendo del punto de inicio, ya que en el agrupamiento de PCA, se observa una mezcla completamente heterogénea de agrupamientos.

En la **Figura 3.26** se muestran los resultados usando el método de reducción de dimensiones t-SNE

El método de agrupamiento K-means con $k = 2$ sobre los datos reducidos en dimensión con t-SNE encuentra un individuo reconocido como valor atípico que, por probabilidad, es un valor vecino de los 4 grupos mayores de vecinos estocásticos. Esto sugiere que existe mayor cercanía

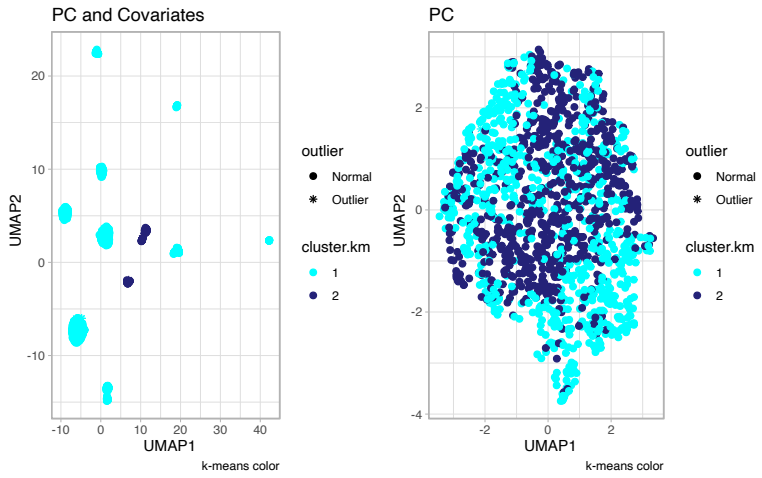

Figura 3.25: Agrupamiento no supervisado aplicando el método K-means de k=2. Método de reducción de dimensiones UMAP aplicado a las componentes principales y covariables (izquierda) y solo componentes principales (derecha).

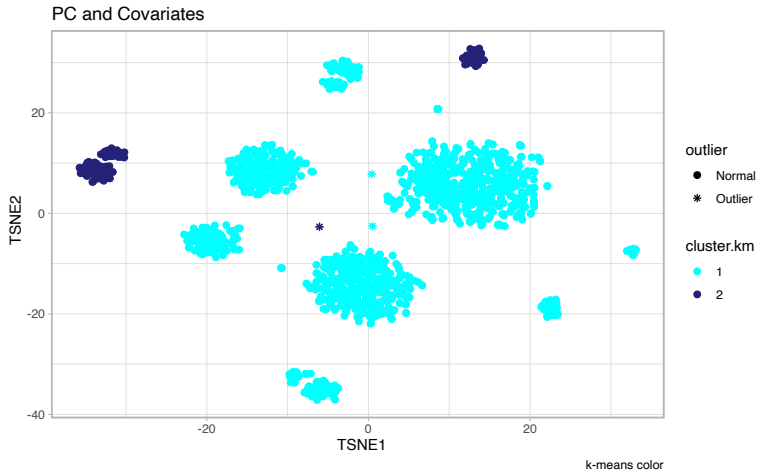

Figura 3.26: Agrupamiento no supervisado aplicando el método K-means de $k = 2$. Método de reducción de dimensiones t-SNE.

de los valores atípicos al agrupamiento 1.

El **Cuadro 3.5** recoge los valores de las muestras que forman parte del agrupamiento 1 y que son considerados como valores atípicos.

A continuación, se ha procedido a la aplicación del algoritmo de agrupamiento no supervisado DBSCAN sobre los datos. Los resultados obtenidos se muestran en la **Figura 3.27** y **Figura 3.28**, aplicados sobre los datos de entrada utilizando una reducción previa de las dimensiones del conjunto de datos mediante UMAP y t-SNE, respectivamente.

En la **Figura 3.28** se observan cinco agrupamientos diferentes, siendo el agrupamiento 1

Paciente	Sexo	Edad	Centro	Mortalidad
196	2	79	3	1
276	2	93	4	2

Cuadro 3.5: Datos de los pacientes considerados como valores atípicos correspondientes al agrupamiento 1 tras aplicar el método K-means con k = 2.

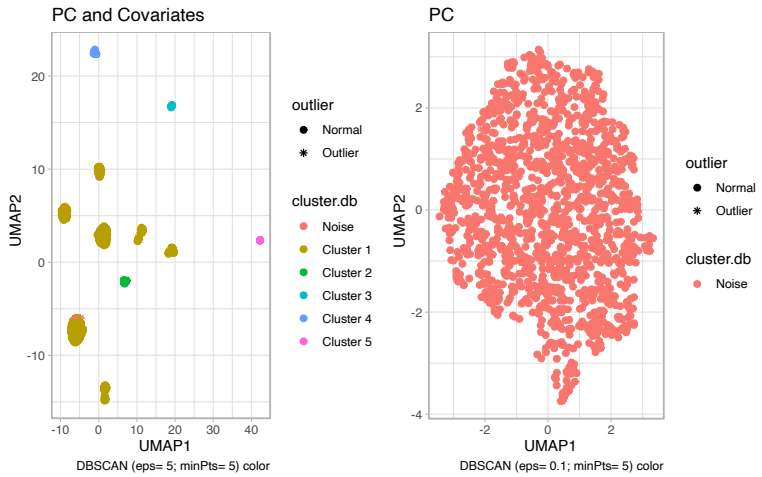

Figura 3.27: Agrupamiento no supervisado aplicando el método DBSCAN. Método de reducción de dimensiones UMAP.

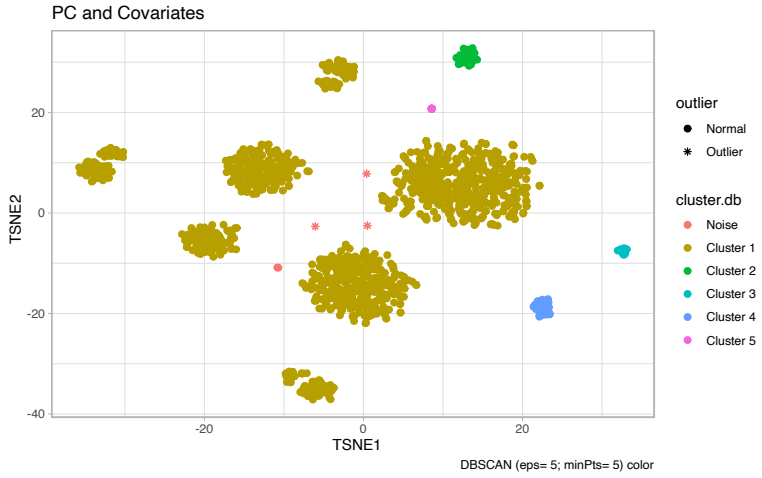

Figura 3.28: Agrupamiento no supervisado aplicando el método DBSCAN. Método de reducción de dimensiones t-SNE.

el que comprende mayor número de individuos. Sin embargo, debido a la elección trivial del parámetro ε, los agrupamientos encontrados pueden variar. Por tanto, la interpretación de este agrupamiento puede resultar comprometida en función del criterio elegido para fijar el parámetro ε.

Debido a ese criterio de elección del parámetro se decide aplicar el método de agrupamiento

no supervisado HDBSCAN, cuyo resultado se muestra en la **Figura 3.29**.

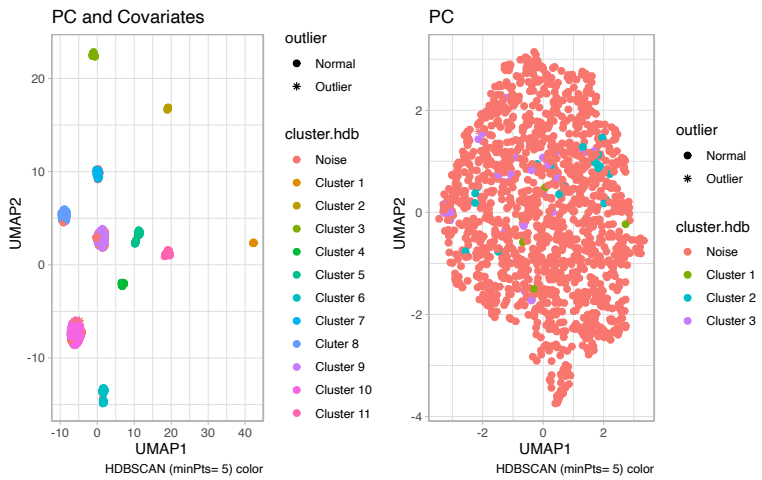

Figura 3.29: Agrupamiento no supervisado aplicando el método HDBSCAN. Método de reducción de dimensiones UMAP.

En la figura se observa que la introducción de covariables al estudio de los datos permite la obtención de más agrupamientos, además estos agrupamientos coinciden con los encontrados al realizar la reducción de dimensiones por el método UMAP.

En el estudio realizado con las componentes principales considera ruido a la mayor parte de los individuos. Sin embargo, HDBSCAN logra encontrar pequeños agrupamientos no observables en la reducción de dimensiones.

El método de agrupamiento no supervisado basado en HDBSCAN permite encontrar un número máximo de agrupamientos según el valor de definido. La **Figura 3.30** muestra la jerarquía de estos agrupamientos en función del valor de utilizado en cada caso.

De arriba a abajo en la **Figura 3.30**, para un intervalo $8 \leq \varepsilon \leq 10$, se identifican dos agrupamientos. En uno de ellos el grupo de individuos está formado por los definidos como *Cluster* 1 en la **Figura 3.29**. Al definir el valor de epsilon $6 \leq \varepsilon \leq 8$ se encuentran 3 agrupamientos y así sucesivamente hasta encontrar los once agrupamientos identificados.

3.4. Herramientas de análisis

Finalmente, como resultado del TFM, también se comparten una serie de herramientas que facilitan la reproducibilidad de los resultados presentados.

Todos los métodos descritos en el presente TFM se encuentran públicamente disponibles en el siguiente repositorio de GitHub. En la Figura 50, se muestra la página principal del repositorio.

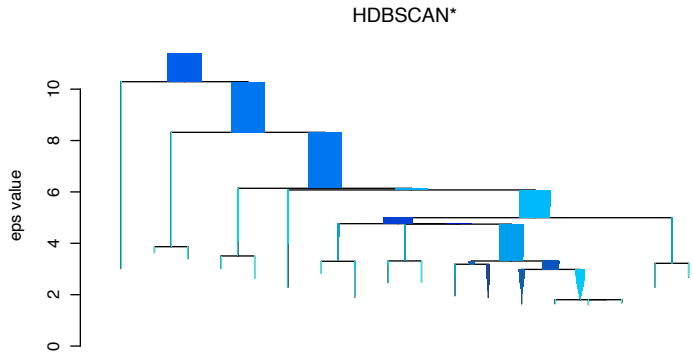

Figura 3.30: Jerarquía de agrupamientos según el valor de epsilon de la matriz de datos genetícos y covariables.

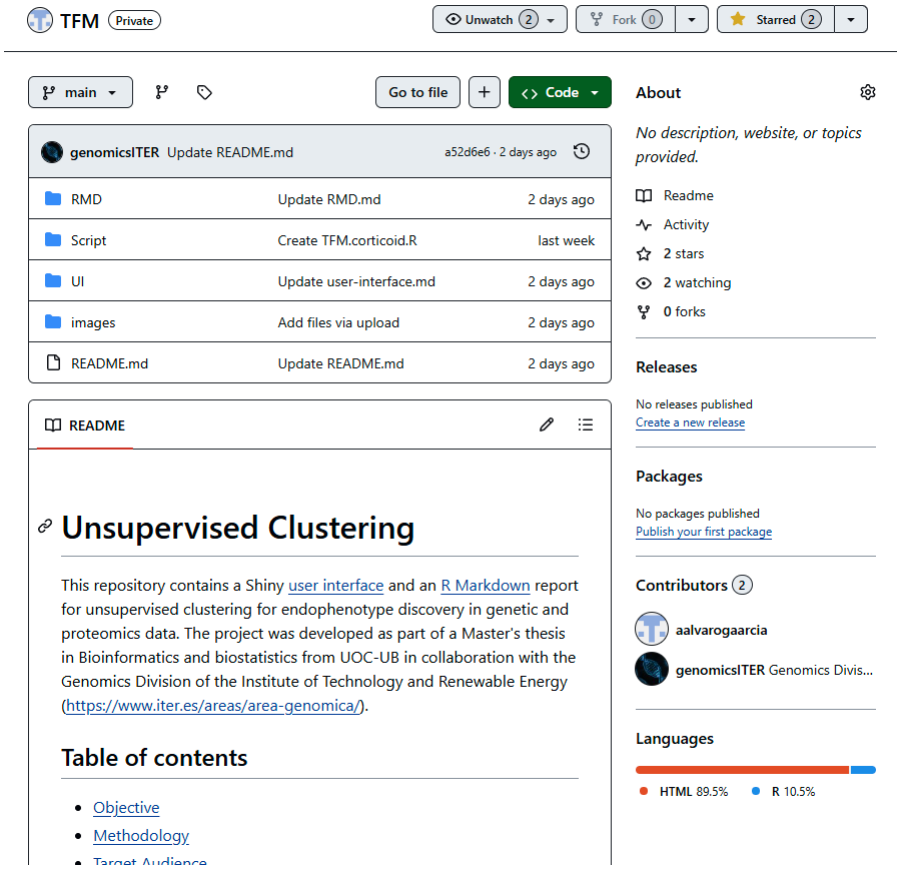

Figura 3.31: Página de inicio del repositorio GitHub del TFM.

En el repositorio se ofrecen tres carpetas, cada una de ellas con un objetivo concreto.

- En la **carpeta RMD** se ofrece un archivo RMarkDown en el que se introducen datos de

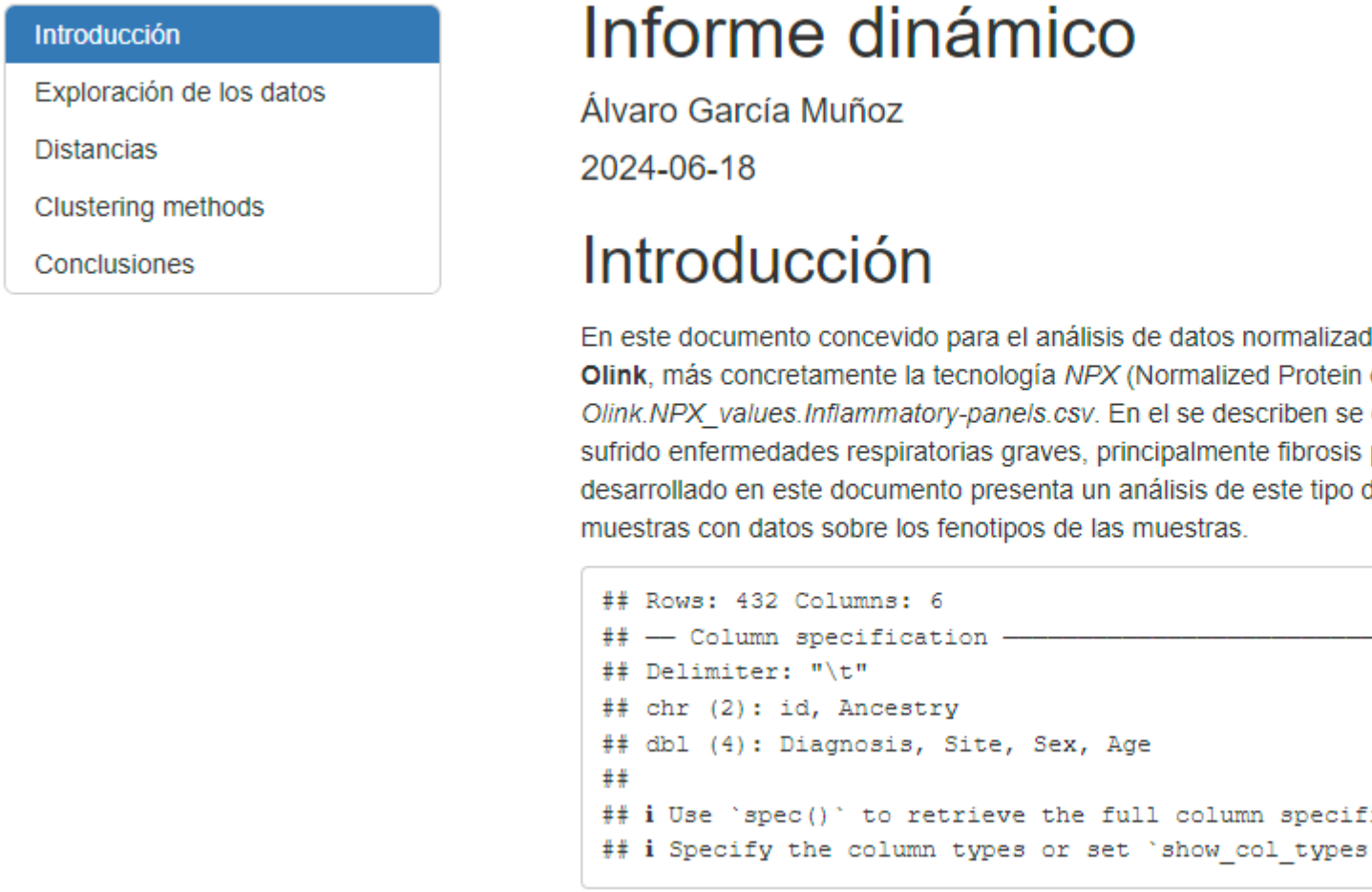

Figura 3.32: Introducción al Informe dinámico en formato HTML, tras la aplicación en datos de proteómica.

proteómica con la misma estructura que la descrita en el **apartado 2.2**. La ejecución del archivo permite obtener un informe dinámico, lo que facilita la replicación del presente estudio (**Figura 3.32**.

- En la **carpeta Script** se ofrece todo el código de R creado y utilizado para el desarrollo de este TFM. Se ofrecen seis scripts con aplicaciones diferentes según el estudio de datos de proteómica de alto rendimiento, datos de genómica en formato VCF o datos procedente del análisis de componentes principales.
- En la **carpeta UI** se encuentra un script desarrollado con la librería shiny que permite la reducción de dimensiones del conjunto de datos de entrada y el agrupamiento no supervisado de datos genotípicos y fenotípicos (**Figura 3.33**).

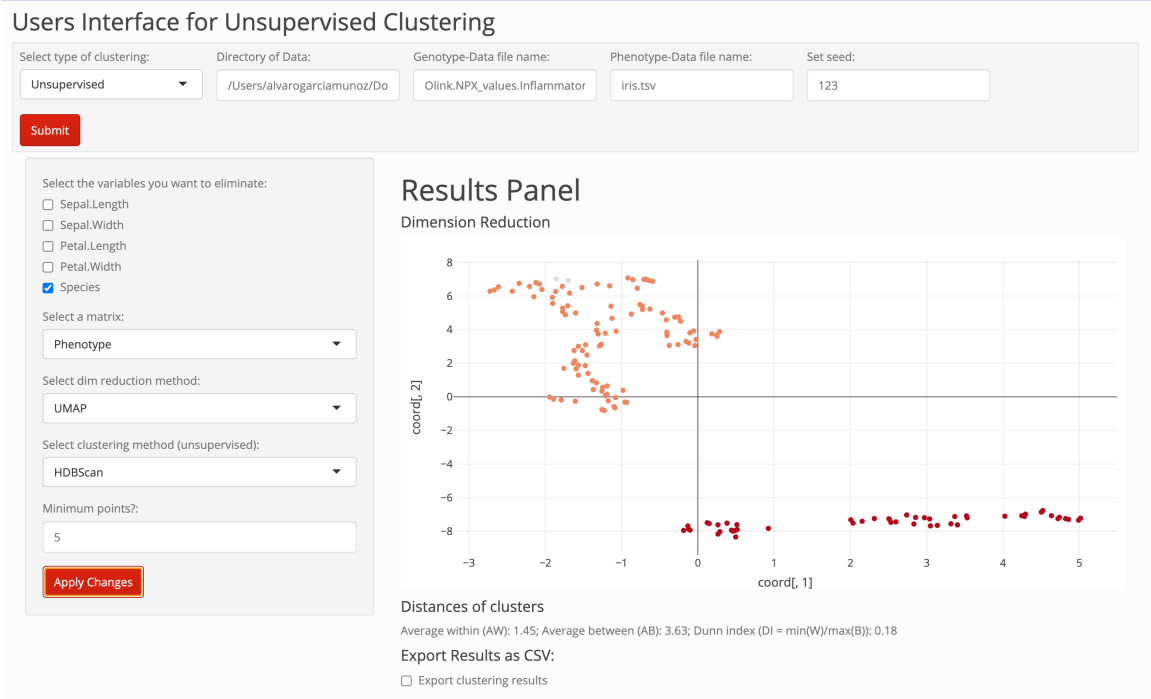

Figura 3.33: Pantalla de análisis de la aplicación Shiny. Muestra los resultados de la reducción de dimensiones y agrupamiento no supervisado del paquete Iris.

Capítulo 4

Conclusiones y trabajos futuros.

4.1. Conclusión

Al comienzo de este trabajo, se plantearon 3 objetivos generales divididos en 6 objetivos específicos. Una vez desarrollado el trabajo en totalidad se extraen las siguientes conclusiones:

- Respecto al objetivo 1, se ha realizado un análisis bibliográfico de técnicas de agrupamiento, y métodos de reducción de dimensiones. Se ha denotado un interés creciente en la identificación de endofenotipos y biomarcadores que faciliten la transición hacia una medicina personalizada o medicina de precisión a partir del análisis de datos ómicos de alto rendimiento. Este interés es especial en el caso de enfermedades con difícil diagnóstico y/o pronóstico, como las enfermedades pulmonares graves.
- Respecto al objetivo 1, se ha reconocido en las recientes investigaciones la aplicación de técnicas y algoritmos de aprendizaje automático o *machine learning* está desempeñando un papel clave en la búsqueda e identificación de endofenotipos y biomarcadores en distintos planos ómicos (genómico, transcriptómico, proteómico, etc.).
- Respecto al objetivo 2, se han diseñado varios productos en el marco de este TFM que permiten reproducir todo el flujo de trabajo bioinformático: un repositorio público en GitHub con todo el código generado, un archivo para la generación automática de informes dinámicos en RMD y una interfaz de usuario programada en shiny para facilitar la reproducibilidad de los resultados obtenidos y su aplicabilidad en otros estudios similares.
- Finalmente respecto al objetivo 3, El flujo de trabajo bioinformático diseñado en este TFM ha sido aplicado a dos estudios en curso sobre enfermedades pulmonares graves, como son la EPIDs y la COVID-19 grave. Para ello se combinan diferentes etapas de análisis exploratorio de los datos con distintos métodos de reducción dimensional y de agrupamiento no supervisado, tanto lineales como no lineales.
- Del análisis realizado de los datos de proteómica de alto rendimienton no se han podido identificar endofenotipos claros. No obstante, se ha detectado un pequeño agrupamiento en los datos de proteómica de alto rendimiento de pacientes con EPID cuya discusión

requiere de información adicional. Este resultado confirma la complejidad y dificultades inherentes a los fenotipos en estudio.

- Del análisis de los datos de pacientes de COVID-19 grave tratados con corticoides se han podido encontrar varios agrupamientos. Sin embargo, para poder definir endofenotipos es necesario profundizar en los agrupamientos encontrados y buscar relacciones biólogicas entre los individuos pertenecientes a un mismo grupo.

El **Cuadro 4.1**, resume conclusiones sobre la metodología empleada en este trabajo:

4.2. Limitaciones y trabajos futuros

Una de las limitaciones del análisis de los datos de proteómica de alto rendimiento de pacientes con EPID es la potencia estadística, toda vez que el conjunto de individuos y de proteínas es relativamente reducido. Se ha de tener en cuenta las limitaciones económicas que impone este tipo de estudios debido al enorme coste por individuo y por proteína analizada para obtener perfiles de proteómica de alto rendimiento (actualmente, en torno a los 1.000 euros/individuo para pocos paneles de proteínas). No obstante, estudios recientes señalan que si se aumenta la potencia estadística, se pueden obtener resultados prometedores [62].

Respecto a los resultados del agrupamiento no supervisado de los datos de pacientes con COVID-19 grave tratados con corticoides, destaca la estructura jerárquica de los agrupamientos obtenida con la aplicación del método HDBSCAN, mostrado en la **Figura 3.30**. Sin embargo, a la hora de la aplicación de los métodos en datos reales, se han encontrado varias dificultades. Una de las dificultades identificadas es la necesidad de cómputo, puesto que el análisis de datos ómicos de alto rendimiento requiere de equipos dotados de capacidades que superan las de un equipo informático de sobremesa. En particular, esta limitación es de importancia si en vez de incluir las componentes principales de los datos genéticos se emplea toda la información genotípica del microarray empleado o los datos genotípicos resultantes de la imputación, lo que conduciría a un espacio de entrada formado por varios millones de variables. En este estudio, la aplicación de distintos métodos de reducción de dimensiones a las distintas matrices de datos de entrada conduce a resultados dispares en función de si se utiliza una aproximación lineal o no lineal, lo que complica la interpretación.

A pesar de las limitaciones y dificultades detalladas anteriormente, para trabajos futuros se propone lo siguiente, en la medida de las posibilidades:

- Aumentar la potencia estadística del estudio en EPIDs, incorporando más individuos y obteniendo el perfil proteómico de mayor número de proteínas.
- Mejorar la calidad de los resultados de los análisis proteómicos, dado que se trata de resultados de expresión semicuantitativa.
- Evaluar distintos algoritmos de selección de variables para abordar el problema de la multicolinealidad observada en los datos de entrada, en particular añadiendo información

Objetivo general	Objetivo específico	Productos	Conclusiones
1. Revisión bibliográfica	1.a Técnicas de *clustering* no supervisado	Geométricos	Sensible a la semilla y la definición de centroides dificulta la búsqueda de endofenotipos.
		Probabilísticos	Suposiciones de distribuciones de probabilidad multivariante. En este trabajo se ha decidido no asumir esas suposiciones.
		Densidad relativa	Sensible a la parametrización. Gran desempeño en la búsqueda de endofenotipos.
	1.b Técnincas de reducción de dimensiones	PCA	Conserva el espacio original. Pierde mucha información en su representación.
		MDS	Conserva las distancias originales. Pierde la topología original.
		t-SNE	Permite la representación de los individuos en baja dimensión. No conserva la estructura espacial.
		UMAP	Permite la representación de los individuos en baja dimensión. No conserva la estructura espacial.

Cuadro 4.1: Resumen de las conclusiones extraídas del trabajo realizado según los objetivos planteados.

sobre la relevancia biológica de las proteínas (procesos biológicos) y, en caso de estar disponible, la expresión génica relacionada con el proteoma.

- Implementación de técnicas de aprendizaje profundo basado en la aplicación de redes neuronales (en sus distintas versiones, como CNN o RNN) y técnicas de análisis topológico de datos, como el que ofrece el entorno kepler-mapper.
- Actualización y ampliación de la interfaz de usuario incorporando nuevos métodos de agrupamiento supervisado y no supervisado, así como alternativas para la reducción de dimensiones.
- Análisis completo del perfil de proteínas o de genes de los agrupamientos para la definición de los biomarcadores combinando las técnicas descritas en el presente TFM con técnicas basadas en el entrenamiento de redes neuronales.

Capítulo 5

Glosario

- **Autovalor (eigenvalor):** Escalar especial asociado a una matriz. Se define como un número que, al multiplicarlo por un vector no nulo asociado a una matriz, resulta en otro vector que es un múltiplo escalar del vector original.
- **Bola:** En topología y otras áreas de la matemática, una bola se define como el conjunto de todos los puntos en el espacio que se encuentran a una distancia menor o igual que una distancia específica, conocida como radio, de un punto central.
- **Colinealidad:** Fenómeno en el ámbito del análisis estadístico y el aprendizaje automático que ocurre cuando dos o más variables independientes en un modelo de regresión lineal están altamente correlacionadas entre sí.
- **Corticoide:** Grupo de hormonas esteroides producidas por las glándulas suprarrenales y sus análogos sintéticos.
- **COVID-19:** Enfermedad infecciosa causada por el virus SARS-CoV-2. El virus se transmite principalmente de persona a persona a través de pequeñas gotitas respiratorias producidas cuando una persona infectada tose, estornuda o habla.
- **Cluster:** Grupo de elementos o datos similares o relacionados que se agrupan en función de alguna característica o variable común.
- **CSV:** *Comma Separated Values*. Formato de archivo de texto en el cual los campos están separados por comas.
- **Deep learning:** Subárea de la Inteligencia Artificial (IA) que se basa en el uso de redes neuronales artificiales profundas para aprender de grandes cantidades de datos
- **Dummy:** También conocida como variable binaria, indicadora o dicotómica, es una variable artificial que se utiliza en análisis estadísticos y modelos de regresión para representar características o atributos categóricos que solo tienen dos categorías.
- **DBSCAN:** De sus siglas en inglés, *Density-Based Spatial Clustering of Applications with Noise*. Algoritmo de agrupamiento de datos sin supervisión, ampliamente utilizado para identificar grupos de puntos en un espacio multidimensional.

- **EDA:** De sus siglas en inglés, *Exploratory Data Analysis*. Conjunto de técnicas y procedimientos estadísticos y de visualización de datos que se utilizan para analizar y comprender las características principales de un conjunto de datos.
- **Endofenotipo:** Un endofenotipo es un rasgo biológico o de comportamiento medible que está genéticamente relacionado con una enfermedad. Los endofenotipos se consideran marcadores intermedios entre los genes y los síntomas clínicos de la enfermedad.
- **EPI/EPID:** La **Enfermedad Pulmonar Intersticial Difusa (EPID)** es un término general que abarca un grupo de más de 200 enfermedades que afectan al tejido intersticial de los pulmones.
- **ETC (CTD):** Enfermedad del Tejido Conectivo, también conocidas como **Enfermedades Reumáticas Autoinmunes**, son un grupo de trastornos crónicos que afectan principalmente al tejido conectivo del cuerpo.
- **FPI:** Fibrosis Pulmonar Idiopática. Enfermedad pulmonar perteneciente al grupo de enfermedades respiratorias conocidas como **enfermedades pulmonares intersticiales difusas fibrosantes.**
- **Genotipado:** El genotipado es el proceso de laboratorio que determina la composición genética específica de un individuo para un conjunto particular de genes o marcadores genéticos.
- **HDBSCAN:** De sus siglas en inglés, *Hierarchical Density-Based Spatial Clustering of Applications with Noise*. Es un algoritmo de clustering jerárquico robusto y de alto rendimiento que se utiliza para identificar grupos de datos (clústeres) en espacios de alta dimensión.
- **Hot-encoding:** También conocido como **codificación binaria**, es una técnica de preprocesamiento de datos utilizada para convertir variables categóricas en representaciones numéricas que los algoritmos de aprendizaje automático pueden procesar y comprender mejor.
- **HPLC-MS/MS:** *Cromatografía líquida acoplada a espectrometría de masas en tándem (HPLC-MS/MS)* es una técnica analítica poderosa que combina la capacidad de separación de la cromatografía líquida de alta resolución (HPLC) con la alta sensibilidad y selectividad de la espectrometría de masas en tándem (MS/MS).
- **K-means:** Algoritmo de agrupamiento no supervisado ampliamente utilizado en el ámbito del aprendizaje automático.
- **Kurtosis:** Medida estadística que describe la forma de la distribución de probabilidad de una variable aleatoria.
- **Machine learning:** Subcampo de la IA que se enfoca en el desarrollo de sistemas que pueden aprender a partir de datos sin ser explícitamente programados.

- **MD:** *MarkDown*. Lenguaje de marcado ligero que facilita la aplicación de formato a un texto empleando una serie de caracteres de una forma especial.
- **MDS:** *MultiDimensional Scaling*. Conjunto de técnicas estadísticas utilizadas para visualizar la estructura de datos de alta dimensionalidad.
- **Microarray:** También conocido como matriz de ADN, es una herramienta de laboratorio utilizada para analizar simultáneamente la expresión de miles de genes en una muestra de ADN o ARN.
- **MinPoints:** *Minimum Points*, en el contexto de algoritmos de agrupamientos, mínimo número de puntos necesarios para crear un agrupamiento.
- **NGS:** *Next Generation Sequencing*. Término genérico que engloba diferentes métodos y herramientas de secuenciación masiva de ácidos nucleicos.
- **NPX:** *Normalized Protein eXpresion*. Unidad de medida arbitraria utilizada por la tecnología de proteómica Olink para cuantificar la expresión de proteínas.
- **OMS:** Organización Mundial de la Salud.
- **PCA:** *Principal Component Analysis*. Técnica estadística multivariante utilizada para reducir la dimensionalidad de un conjunto de datos.
- **PEA:** *Proximity Extension Assay*. Ensayo de extensión de proximidad. Método para detectar y cuantificar la cantidad de muchas proteínas específicas presentes en una muestra biológica.
- **PLINK:** Conjunto de herramientas de software de código abierto para el análisis genético.
- **Proteómica:** Estudio a gran escala de las concentraciones de proteínas de una muestra.
- **RMD:** *R MarkDown*. Lenguaje basado en MD desarrollado específicamente para el uso de R.
- **Scree-plot:** Diagrama de sedimentación. Gráfica lineal que muestra los autovalores de los componentes principales (PCs) SVM: Support Vector Machines. Tipo de algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión en problemas de Machine Learning.
- **t-SNE:** *t-distributed Stochastic Neighbor Embedding*. Método estadístico de reducción de dimensiones basado en la distribución de probabilidad t de Student.
- **TSV:** *Tabulator Separated Values*. Formato de archivo de texto en el cual los campos están separados por tabuladores.
- **UI:** *Users Interface*. Conjunto de elementos visuales y funcionales que permiten a los usuarios interactuar con un sistema o aplicación digital.

- **UMAP:** *Uniform Manifold Approximations and Projections*. Algoritmo de reducción de dimensiones basado en técnicas de aprendizaje de múltiples e ideas del análisis topológico de datos.
- **VCF:** *Variant Call Format*. Formato de archivo de texto estandarizado que se utiliza en bioinformática para almacenar y compartir información sobre las variantes genéticas.

Capítulo 6

Bibliografía

[1] "How is the olink® target 96 npx data pre-processed? - olink." [Online]. Available: https://olink.com/faq/how-is-the-data-pre-processed/

[2] S. S. King, R. A. Rahman, M. A. Fauzi, and A. T. Haron, "Critical analysis of pandemic impact on aec organizations: the covid-19 case," *Journal of Engineering, Design and Technology*, vol. 20, no. 1, pp. 358–383, 2022.

[3] D. CHEN, H. LONG, S. LI, and Y. CHEN, "Interpretation of global strategy for the diagnosis, treatment, management and prevention of chronic obstructive pulmonary disease 2024 report," *Chinese General Practice*, vol. 27, no. 13, p. 1533, 2024.

[4] M. P. Davies, T. Sato, H. Ashoor, L. Hou, T. Liloglou, R. Yang, and J. K. Field, "Plasma protein biomarkers for early prediction of lung cancer," *eBioMedicine*, vol. 93, p. 104686, 7 2023.

[5] "Endophenotype - an overview — sciencedirect topics." [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/endophenotype

[6] I. Metzler, "Biomarkers and their consequences for the biomedical profession: a social science perspective," *Personalized Medicine*, vol. 7, no. 4, pp. 407–420, 2010.

[7] A. H. Lampezhev, E. Y. Linskaya, A. A. Tatarkanov, and I. A. Alexandrov, "Cluster data analysis with a fuzzy equivalence relation to substantiate a medical diagnosis," *Emerging Science Journal*, vol. 5, no. 5, pp. 688–699, 2021.

[8] W. Chang, J. Cheng, J. J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, "shiny: Web application framework for r," 2024. [Online]. Available: https://CRAN.R-project.org/package=shiny

[9] H. Alashwal, M. E. Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The application of unsupervised clustering methods to alzheimer's disease," *Frontiers in Computational Neuroscience*, vol. 13, 5 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fncom.2019.00031/full

[10] M. R. Kosorok and E. B. Laber, "Precision medicine," *Annual Review of Statistics and Its Application*, vol. 6, pp. 263–286, 3 2019.

[11] Y. Jiang, X. Zhou, F. C. Ip, P. Chan, Y. Chen, N. C. Lai, K. Cheung, R. M. Lo, E. P. Tong, B. W. Wong, A. L. Chan, V. C. Mok, T. C. Kwok, K. Y. Mok, J. Hardy, H. Zetterberg, A. K.

Fu, and N. Y. Ip, "Large-scale plasma proteomic profiling identifies a high-performance biomarker panel for alzheimer's disease screening and staging," *Alzheimer's Dementia*, vol. 18, pp. 88–102, 1 2022.

[12] H. Desaire, E. P. Go, and D. Hua, "Advances, obstacles, and opportunities for machine learning in proteomics," *Cell Reports Physical Science*, vol. 3, p. 101069, 10 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666386422003630

[13] G. Raghu, M. Remy-Jardin, J. L. Myers, L. Richeldi, C. J. Ryerson, D. J. Lederer, J. Behr, V. Cottin, S. K. Danoff, F. Morell, K. R. Flaherty, A. Wells, F. J. Martinez, A. Azuma, T. J. Bice, D. Bouros, K. K. Brown, H. R. Collard, A. Duggal, L. Galvin, Y. Inoue, R. G. Jenkins, T. Johkoh, E. A. Kazerooni, M. Kitaichi, S. L. Knight, G. Mansour, A. G. Nicholson, S. N. J. Pipavath, I. Buendía-Roldán, M. Selman, W. D. Travis, S. L. F. Walsh, and K. C. Wilson, "Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline," *American Journal of Respiratory and Critical Care Medicine*, vol. 198, pp. e44–e68, 9 2018. [Online]. Available: https://www.atsjournals.org/doi/10.1164/rccm.201807-1255ST

[14] Y. Huang, S.-F. Ma, J. M. Oldham, A. Adegunsoye, D. Zhu, S. Murray, J. S. Kim, C. Bonham, E. Strickland, A. L. Linderholm, C. T. Lee, T. Paul, H. Mannem, T. M. Maher, P. L. Molyneaux, M. E. Strek, F. J. Martinez, and I. Noth, "Machine learning of plasma proteomics classifies diagnosis of interstitial lung disease," *American Journal of Respiratory and Critical Care Medicine*, 2 2024.

[15] L. M. Kraven, A. R. Taylor, P. L. Molyneaux, T. M. Maher, J. E. McDonough, M. Mura, I. V. Yang, D. A. Schwartz, Y. Huang, I. Noth, S. F. Ma, A. J. Yeo, W. A. Fahy, R. G. Jenkins, and L. V. Wain, "Cluster analysis of transcriptomic datasets to identify endotypes of idiopathic pulmonary fibrosis," *Thorax*, vol. 78, p. 551, 6 2023. [Online]. Available: http://thorax.bmj.com/content/78/6/551.abstract

[16] L. Winchester, I. Barber, M. Lawton, J. Ash, B. Liu, S. Evetts, L. Hopkins-Jones, S. Lewis, C. Bresner, A. B. Malpartida, N. Williams, S. Gentlemen, R. Wade-Martins, B. Ryan, A. Holgado-Nevado, M. Hu, Y. Ben-Shlomo, D. Grosset, and S. Lovestone, "Identification of a possible proteomic biomarker in parkinson's disease: discovery and replication in blood, brain and cerebrospinal fluid," *Brain Communications*, vol. 5, 12 2022.

[17] N. B. Palstrøm, R. Matthiesen, L. M. Rasmussen, and H. C. Beck, "Recent developments in clinical plasma proteomics—applied to cardiovascular research," *Biomedicines*, vol. 10, p. 162, 1 2022.

[18] "Data — 1000 genomes." [Online]. Available: https://www.internationalgenome.org/data

[19] "Genomics data - iter - instituto tecnológico y de energías renovables, s.a." [Online]. Available: https://www.iter.es/portfolio-items/datos/?portfolioCats=367

[20] "Scourge covid." [Online]. Available: https://www.scourge-covid.org/

[21] V. Cottin, N. A. Hirani, D. L. Hotchkin, A. M. Nambiar, T. Ogura, M. Otaola, D. Skowasch, J. S. Park, H. K. Poonyagariyagorn, W. Wuyts, and A. U. Wells, "Presentation, diagnosis and clinical course of the spectrum of progressive-fibrosing interstitial lung diseases," *European Respiratory Review*, vol. 27, p. 180076, 12 2018. [Online]. Available: http://err.ersjournals.com/lookup/doi/10.1183/16000617.0076-2018

[22] R. Cruz, S. D. de Almeida, M. L. de Heredia, I. Quintela, F. C. Ceballos, G. Pita, J. M. Lorenzo-Salazar, R. González-Montelongo, M. Gago-Domínguez, M. S. Porras, J. A. T. Castaño, J. Nevado, J. M. Aguado, C. Aguilar, S. Aguilera-Albesa, V. Almadana, B. Almoguera, N. Alvarez, Álvaro Andreu-Bernabeu, E. Arana-Arri, C. Arango, M. J. Arranz, M.-J. Artiga, R. C. Baptista-Rosas, M. Barreda-Sánchez, M. Belhassen-Garcia, J. F. Bezerra, M. A. C. Bezerra, L. Boix-Palop, M. Brion, R. Brugada, M. Bustos, E. J. Calderón, C. Carbonell, L. Castano, J. E. Castelao, R. Conde-Vicente, M. L. Cordero-Lorenzana, J. L. Cortes-Sanchez, M. Corton, M. T. Darnaude, A. D. Martino-Rodríguez, V. del Campo-Pérez, A. D. de Bustamante, E. Domínguez-Garrido, A. D. Luchessi, R. Eiros, G. M. E. Sanabria, M. C. Fariñas, U. Fernández-Robelo, A. Fernández-Rodríguez, T. Fernández-Villa, B. Gil-Fournier, J. Gómez-Arrue, B. G. Álvarez, F. G. B. de Quirós, J. González-Peñas, J. F. Gutiérrez-Bautista, M. J. Herrero, A. Herrero-Gonzalez, M. A. Jimenez-Sousa, M. C. Lattig, A. L. Borja, R. Lopez-Rodriguez, E. Mancebo, C. Martín-López, V. Martín, O. Martinez-Nieto, I. Martinez-Lopez, M. F. Martinez-Resendez, A. Martinez-Perez, J. F. Mazzeu, E. M. Macías, P. Minguez, V. M. Cuerda, V. N. Silbiger, S. F. Oliveira, E. Ortega-Paino, M. Parellada, E. Paz-Artal, N. P. C. Santos, P. Pérez-Matute, P. Perez, M. E. Pérez-Tomás, T. Perucho, M. L. Pinsach-Abuin, E. N. Pompa-Mera, G. L. Porras-Hurtado, A. Pujol, S. R. León, S. Resino, M. R. Fernandes, E. Rodríguez-Ruiz, F. Rodriguez-Artalejo, J. A. Rodriguez-Garcia, F. R. Cabello, J. Ruiz-Hornillos, P. Ryan, J. M. Soria, J. C. Souto, E. Tamayo, A. Tamayo-Velasco, J. C. Taracido-Fernandez, A. Teper, L. Torres-Tobar, M. Urioste, J. Valencia-Ramos, Z. Yáñez, R. Zarate, T. Nakanishi, S. Pigazzini, F. Degenhardt, G. Butler-Laporte, D. Maya-Miles, L. Bujanda, Y. Bouysran, A. Palom, D. Ellinghaus, M. Martínez-Bueno, S. Rolker, S. Amitrano, L. Roade, F. Fava, C. D. Spinner, D. Prati, D. Bernardo, F. Garcia, G. Darcis, I. Fernández-Cadenas, J. C. Holter, J. M. Banales, R. Frithiof, S. Duga, R. Asselta, A. C. Pereira, M. Romero-Gómez, B. Nafría-Jiménez, J. R. Hov, I. Migeotte, A. Renieri, A. M. Planas, K. U. Ludwig, M. Buti, S. Rahmouni, M. E. Alarcón-Riquelme, E. C. Schulte, A. Franke, T. H. Karlsen, L. Valenti, H. Zeberg, B. Richards, A. Ganna, M. Boada, I. de Rojas, A. Ruiz, P. Sánchez-Juan, L. M. Real, E. Guillen-Navarro, C. Ayuso, A. González-Neira, J. A. Riancho, A. Rojas-Martinez, C. Flores, P. Lapunzina, and A. Carracedo, "Novel genes and sex differences in covid-19 severity," *Human Molecular Genetics*, vol. 31, pp. 3789–3806, 11 2022.

[23] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[24] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002.

[25] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2 2018.

[26] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante, "Genes mirror geography within europe," *Nature*, vol. 456, pp. 98–101, 11 2008.

[27] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, pp. 48–50, 1956.

[28] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 3 1951.

[29] H.-P. Kriegel, E. Schubert, and A. Zimek, "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?" *Knowledge and Information Systems*, vol. 52, pp. 341–378, 8 2017.

[30] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited," *ACM Transactions on Database Systems*, vol. 42, pp. 1–21, 9 2017.

[31] R. J. G. B. Campello, D. Moulavi, and J. Sander, *Density-Based Clustering Based on Hierarchical Density Estimates*, 2013, pp. 160–172.

[32] R. C. Team, "R: A language and environment for statistical computing," 2024. [Online]. Available: https://www.R-project.org/

[33] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[34] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 2010, pp. 51–56.

[35] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[36] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[37] R. Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/

[38] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, p. 1686, 11 2019.

[39] M. Ballings and D. Van den Poel, *dummy: Automatic Creation of Dummies with Support for Predictive Modeling*, 2015, r package version 0.1.3. [Online]. Available: https://CRAN.R-project.org/package=dummy

[40] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: https://www.stats.ox.ac.uk/pub/MASS4/

[41] X. Zheng, D. Levine, J. Shen, S. Gogarten, C. Laurie, and B. Weir, "A high-performance computing toolset for relatedness and principal component analysis of snp data," *Bioinformatics*, vol. 28, no. 24, pp. 3326–3328, 2012.

[42] T. Konopka, *umap: Uniform Manifold Approximation and Projection*, 2023, r package version 0.2.10.0. [Online]. Available: https://CRAN.R-project.org/package=umap

[43] J. H. Krijthe, *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*, 2015, r package version 0.17. [Online]. Available: https://github.com/jkrijthe/Rtsne

[44] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[45] L. van der Maaten, "Accelerating t-sne using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, pp. 3221–3245, 2014.

[46] M. Hahsler and M. Piekenbrock, *dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, 2023, r package version 1.1-12. [Online]. Available: https://CRAN.R-project.org/package=dbscan

[47] M. Hahsler, M. Piekenbrock, and D. Doran, "dbscan: Fast density-based clustering with R," *Journal of Statistical Software*, vol. 91, no. 1, pp. 1–30, 2019.

[48] Wilkerson, M. D., Hayes, and D. Neil, "Consensusclusterplus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, pp. 1572–1573, 2010. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/26/12/1572.abstract

[49] W. Chang, *shinythemes: Themes for Shiny*, 2021, r package version 1.2.0. [Online]. Available: https://CRAN.R-project.org/package=shinythemes

[50] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[51] C. Sievert, *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. [Online]. Available: https://plotly-r.com

[52] B. Auguie, *gridExtra: Miscellaneous Functions for "Grid"Graphics*, 2017, r package version 2.3. [Online]. Available: https://CRAN.R-project.org/package=gridExtra

[53] "La oms revela las principales causas de muerte y discapacidad en el mundo: 2000-2019." [Online]. Available: https://www.who.int/es/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019

[54] "Report of the sixteenth annual meeting of the global alliance against chronic respiratory diseases: virtual meeting, 12 december 2023." [Online]. Available: https://www.who.int/publications/i/item/9789240089822

[55] "Definition of biomarker - nci dictionary of cancer terms - nci." [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker

[56] J. Molina, J. Trigueros, J. Quintano, E. Mascarós, A. Xaubet, and J. Ancochea, "Fibrosis pulmonar idiopática: un reto para la atención primaria," *SEMERGEN - Medicina de Familia*, vol. 40, pp. 134–142, 4 2014.

[57] J. M. Oldham, Y. Huang, S. Bose, S.-F. Ma, J. S. Kim, A. Schwab, C. Ting, K. Mou, C. T. Lee, A. Adegunsoye, S. Ghodrati, J. V. Pugashetti, N. Nazemi, M. E. Strek, A. L. Linderholm, C.-H. Chen, S. Murray, R. L. Zemans, K. R. Flaherty, F. J. Martinez, and I. Noth, "Proteomic biomarkers of survival in idiopathic pulmonary fibrosis," *American Journal of Respiratory and Critical Care Medicine*, vol. 209, pp. 1111–1120, 5 2024.

[58] F. Liew, C. Efstathiou, S. Fontanella, M. Richardson, R. Saunders, D. Swieboda, J. K. Sidhu, S. Ascough, S. C. Moore, N. Mohamed, J. Nunag, C. King, O. C. Leavy, O. Elneima, H. J. C. McAuley, A. Shikotra, A. Singapuri, M. Sereno, V. C. Harris, L. Houchen-Wolloff, N. J. Greening, N. I. Lone, M. Thorpe, A. A. R. Thompson, S. L. Rowland-Jones, A. B. Docherty, J. D. Chalmers, and L.-P. H. et al., "Large-scale phenotyping of patients with long covid post-hospitalization reveals mechanistic subtypes of disease," *Nature Immunology*, vol. 25, pp. 607–621, 4 2024.

[59] L. Lavagnino, F. Amianto, B. Mwangi, F. D'Agata, A. Spalatro, G. B. Zunta-Soares, G. A. Daga, P. Mortara, S. Fassino, and J. C. Soares, "Identifying neuroanatomical signatures of anorexia nervosa: a multivariate machine learning approach," *Psychological Medicine*, vol. 45, pp. 2805–2812, 10 2015. [Online]. Available: https://www.cambridge.org/core/journals/psychological-medicine/article/abs/identifying-neuroanatomical-signatures-of-anorexia-nervosa-a-multivariate-machine-learning-approach/372CA0F2FD09985001B2A9D12FDE0E83

[60] J. Hair, *Multivariate data analysis*, 2009.

[61] W. S. Bowman, C. A. Newton, A. L. Linderholm, M. L. Neely, J. V. Pugashetti, B. Kaul, V. Vo, G. A. Echt, W. Leon, R. J. Shah, Y. Huang, C. K. Garcia, P. J. Wolters, and J. M. Oldham, "Proteomic biomarkers of progressive fibrosing interstitial lung disease: a multicentre cohort analysis," *The Lancet Respiratory Medicine*, vol. 10, pp. 593–602, 6 2022.

[62] A. M. W. Lim, E. U. Lim, P.-L. Chen, and C. S. J. Fann, "Cluster analysis identified clinically relevant metabolic syndrome endophenotypes," *medRxiv*, p. 2022.11.04.22281926, 1 2022. [Online]. Available: http://medrxiv.org/content/early/2022/11/05/2022.11.04.22281926.abstract

Apéndice A

Figuras suplementarias

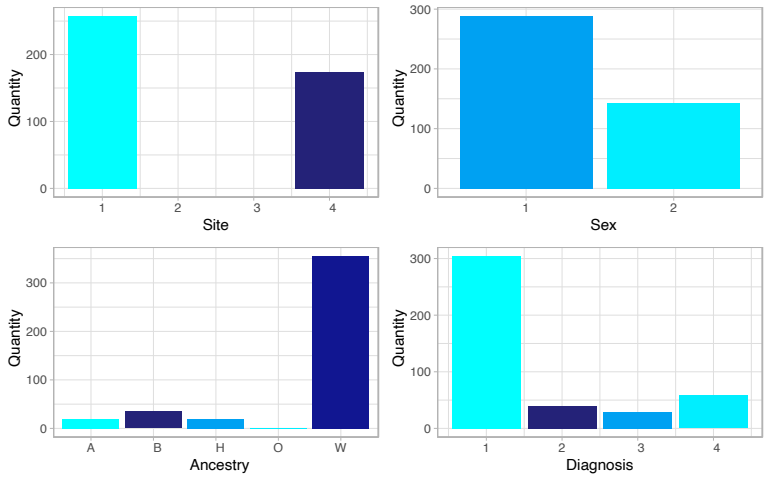

Figura A.1: Diagrama de barras de 3 variables discretas y la variable fenotípica.

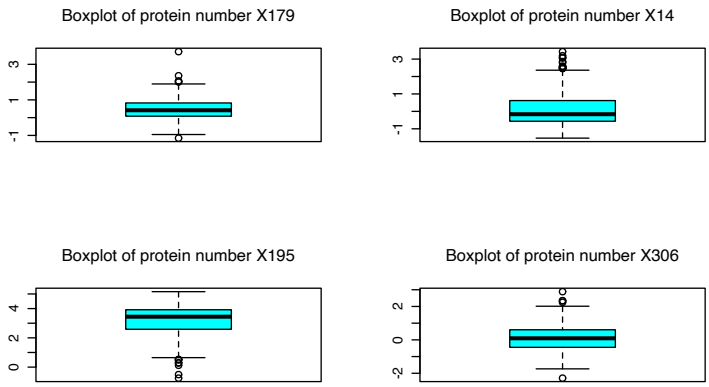

Figura A.2: Diagrama de cajas de las proteínas 179, 14, 195 y 306, escogidas aleatoriamente entre el total de proteínas.

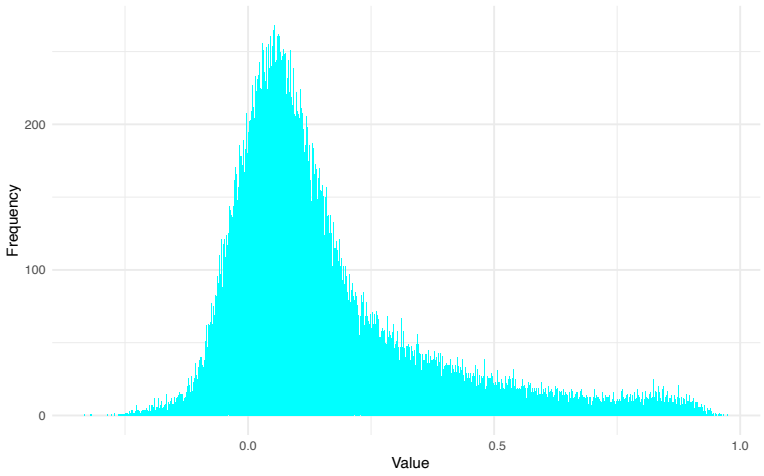

Figura A.3: Histograma de la correlación de Pearson de las proteínas de las muestras de pacientes de EPID.

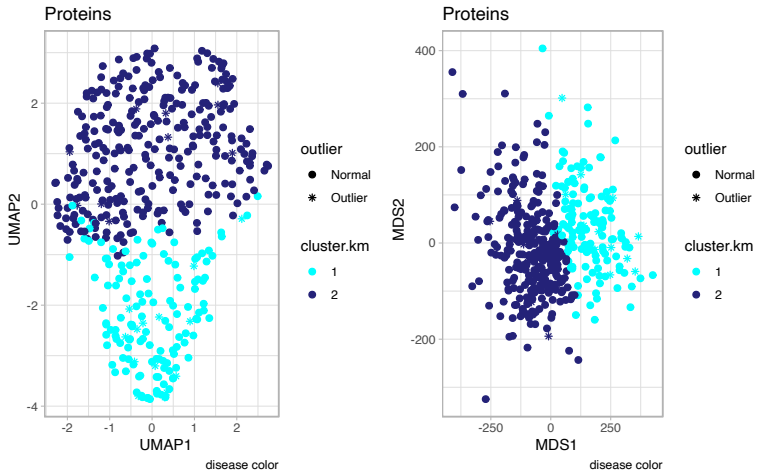

Figura A.4: Representación de los datos de proteínas con el método de reducción de dimensiones de UMAP coloreado según el diagnóstico (izquierda) y según el método de agrupamiento no supervisado (derecha).

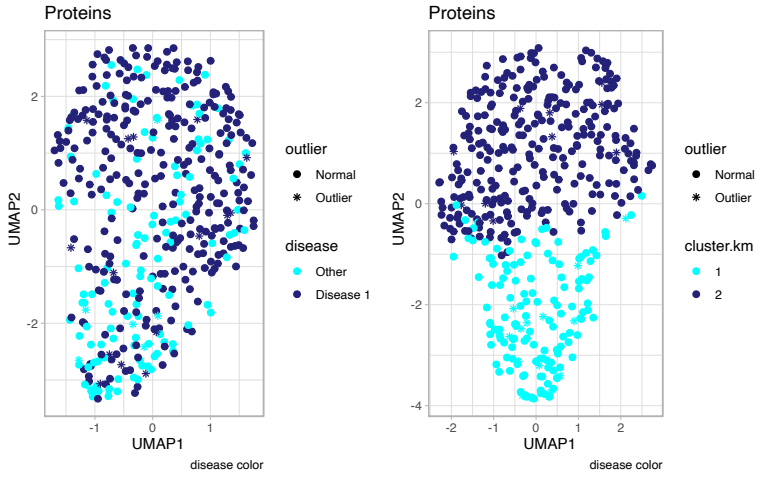

Figura A.5: Representación de los datos de proteínas con el método de reducción de dimensiones de UMAP coloreado según el diagnóstico (izquierda) y según el método de agrupamiento no supervisado (derecha).

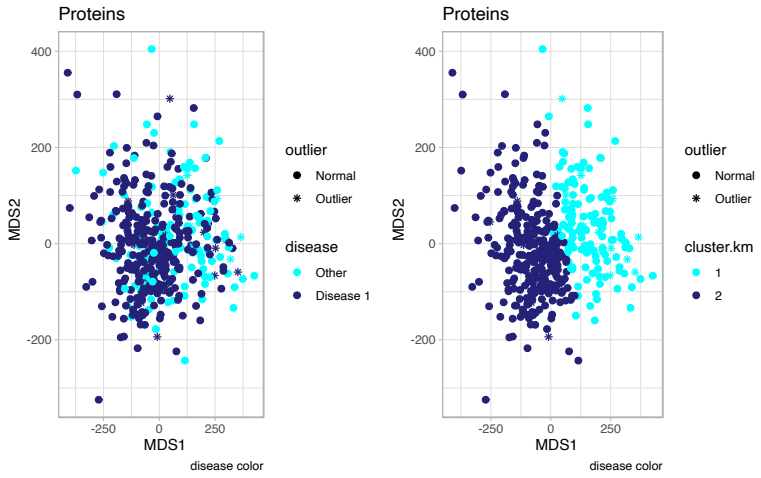

Figura A.6: Representación de los datos de proteínas con el método de reducción de dimensiones de MDS coloreado según el diagnóstico (izquierda) y según el método de agrupamiento no supervisado (derecha).

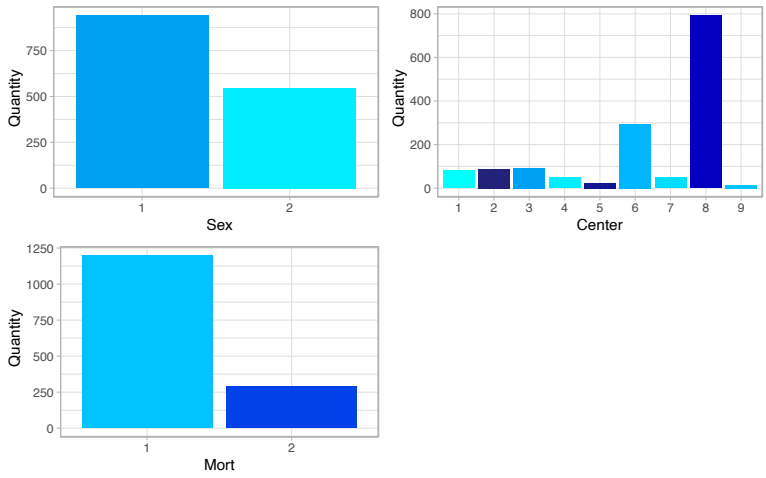

Figura A.7: Distribución de las variables discretas de los pacientes con COVID-19 grave tratados con corticoides.

Apéndice B

Kurtosis y Sesgo

Los estadísticos de sesgo y kurtosis son medidas importantes que describen la forma de la distribución de una variable aleatoria. Estas medidas proporcionan información valiosa sobre la simetría y la concentración de los datos alrededor de la media.

El sesgo es una medida de la asimetría de una distribución. Una distribución simétrica tiene un sesgo de cero, mientras que una distribución asimétrica tiene un sesgo positivo o negativo, dependiendo de la dirección de la asimetría.

La kurtosis es una medida de la concentración de los datos alrededor de la media. Una distribución con kurtosis normal tiene una forma de campana, mientras que una distribución con kurtosis alta tiene picos más altos y colas más pesadas que una distribución normal. Una distribución con kurtosis baja tiene una forma más plana que una distribución normal.

La kurtosis se puede calcular utilizando diferentes medidas, como la kurtosis muestral o la kurtosis poblacional. La kurtosis muestral se calcula a partir de una muestra de datos, mientras que la kurtosis poblacional se calcula a partir de toda la población.

Los estadísticos de sesgo y kurtosis son herramientas valiosas para describir la forma de una distribución de datos. La comprensión de estas medidas puede ser útil para una variedad de aplicaciones en análisis de datos, modelado estadístico e inferencia estadística.

Apéndice C

Distancia de Mahalanobbis

La distancia de Mahalanobis, introducida por el estadístico indio Prasanta Chandra Mahalanobis en 1936, es una medida de distancia multivariante que se utiliza para determinar la similitud entre dos vectores de datos. A diferencia de la distancia euclidiana, que solo considera la distancia física entre dos puntos, la distancia de Mahalanobis toma en cuenta la correlación entre las variables y la varianza de cada una de ellas.

La distancia de Mahalanobis entre dos puntos x e y, se define como:

$$d_{Mahalanobis} = \sqrt{(x-y)' \times \Sigma^{-1}(x-y)} \tag{C.1}$$

donde Σ es la matriz de covarianza del vector aleatorio $X := \{x_1, ..., x_n\}$ y $(x-y)'$ es la transpuesta del vector $(x-y)$.

Para la detección de valores atípicos multivariante, se ha aplicado la distancia del conjunto de individuos $M := \{m_1, ..., m_n\}$ y el punto medio μ robusto.

Se introduce el código aquí:

```
    #Robust Mean point
mu <- c()
for (i in 2:ncol(proteins)){
  temp <- rlm(proteins[,i]~1, maxit = 100)$coefficients
  mu <- cbind(mu ,temp)
}

    #Robust covariance matrix
sig <- cov.rob(proteins[,-1])$cov

    #Distance to mu (Mahalanobis)
dist.mu <- apply(proteins[,-1], 1, function(x) mahalanobis(x, mu, sig))
dist.mu <- data.frame(Index = 1:length(dist.mu), distance = dist.mu)
```

Apéndice D

Aplicación Shiny

[h!] El presente anexo tiene como objetivo describir la UI desarrollada para el TFM, la cual facilita el procesamiento de agrupamiento no supervisado de datos.

Para comprender el funcionamiento, se empieza por analizar la composición de la interfaz (**Figura D.1**).

La interfaz de usuario se ha diseñado por *bloques*, como se observa en la **Figura D.2**.

D.1. Bloque de datos

Es el bloque superior de la interfaz, señalada en la **Figura D.3**, en el se selecciona el tipo de agrupamiento que se desea realizar. Por el momento solo se encuentra disponibles las opciones *Unsupervised* (No supervisado) y *None* (Sin agrupamiento). A continuación, se ha de introducir el directorio del ordenador donde se encuentran los archivos de datos que se desean estudiar, así como los nombres de los archivos.

Se ha de tener en cuenta que los archivos han de haber sido procesados previamente. Para el agrupamiento no supervisado se ha de introducir un archivo genotípico en formato CSV sin identificadores, ha de tratarse de una matriz enteramente numérica. De manera análoga, se ha de introducir un archivo con información de las covariables en formato TSV sin identificadores, ha de tratarse de una matriz enteramente numérica.

D.2. Bloque de agrupamiento

En el bloque izquierdo de la interfaz de usuario se encuentran las opciones para realizar el agrupamiento no supervisado. En él se seleccionan los datos que se desean analizar y el procedimiento del análisis.

Se selecciona entre las tres matrices explicadas a lo largo de este trabajo. Posteriormente, se selecciona un método de reducción de dimensiones y las variable del archivo *phenotypes* que

se desea definir como *target*. Una vez se ha seleccionado el modo de visualización, se procede a escoger el método de agrupamiento que se desea aplicar.

D.3. Bloque de resultados

En el bloque de resultados se encuentra, principalmente, una gráfica interactiva en la que se observa las muestras en dos dimensiones coloreadas según su agrupamiento.

Además, en la parte inferior se ofrecen tres medidas de comparación de agrupamientos no supervisados.

- *Average Within (AW):* Media de distancias en los agrupamientos
- *Averege Between (AB):* Media de distancia entre agrupamientos.
- *Dunn Index (DI):* La proporción entre la distancia minima en el agrupamiento entre la distancia máxima entre agrupamientos.

Finalmente, se ofrece una opción que permite exportar los resultados en formato CSV.

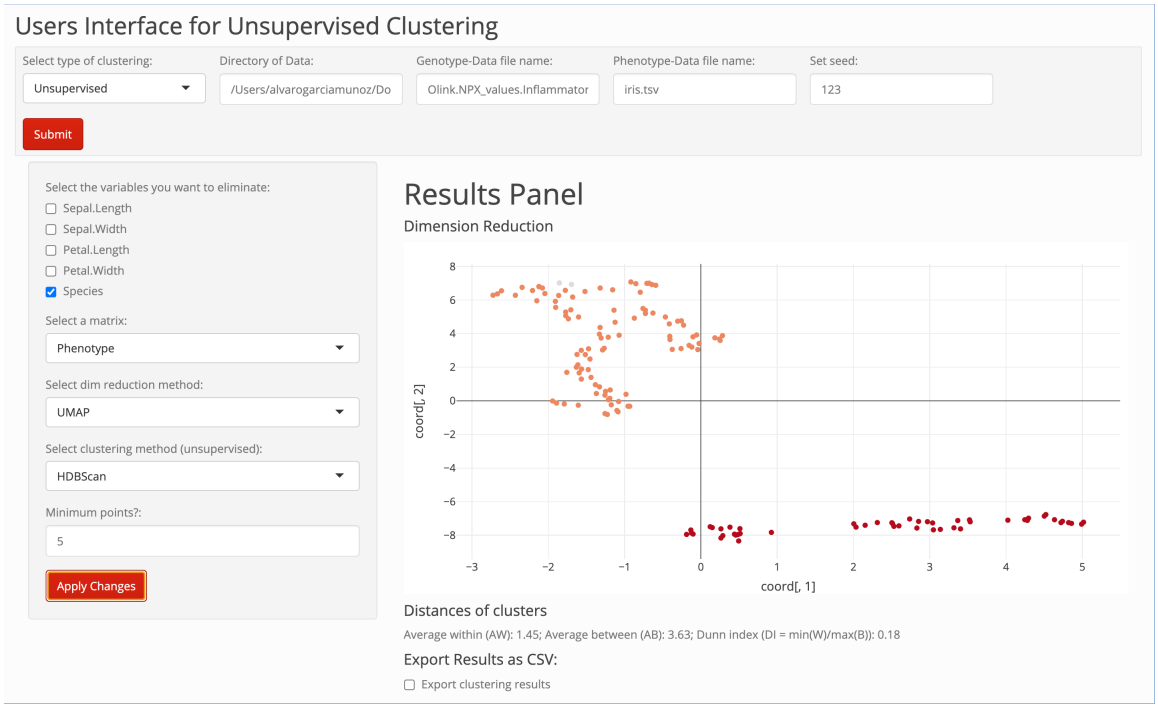

Figura D.1: Interfaz de usuario desarrollada en shiny.

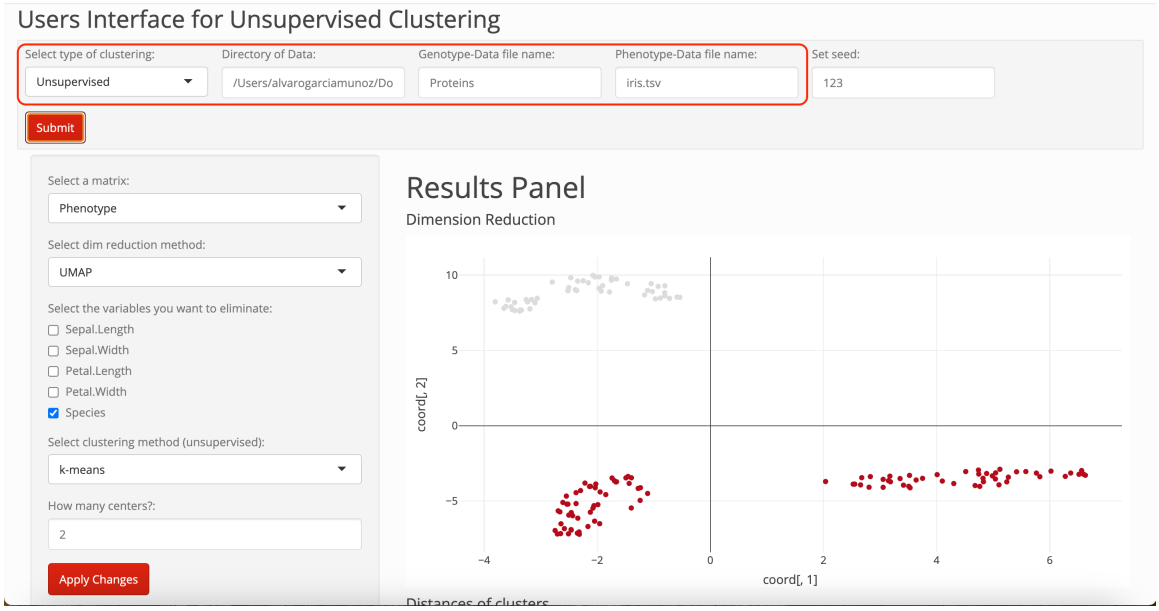

Figura D.2: Bloque de datos, de izquierda a derecha se ha de señalar: tipo de agrupamiento, directorio de datos, datos genotípicos, datos variables y semilla.

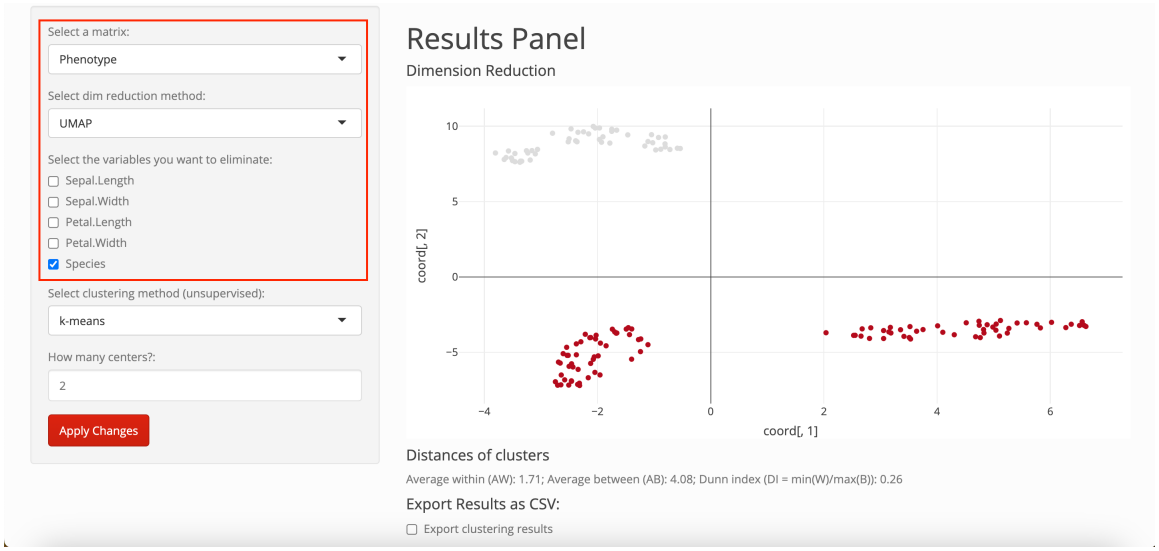

Figura D.3: Bloque de agrupamiento, de arriba a abajo se ha de seleccionar: matriz de análisis, método de reducción de dimensiones, variables a obviar para el analisis, método de agrupamiento no supervisado y parámetros.