



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa
Aplicadas y Calidad

Clasificación de pacientes de melanoma en grupos
etiopatogénicos

Trabajo Fin de Máster

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

AUTOR/A: Aguerralde Martin, Maider

Tutor/a: Tarazona Campos, Sonia

Cotutor/a externo: NAGORE ENGUIDANOS, EDUARDO

CURSO ACADÉMICO: 2022/2023

Agradecimientos

Quiero agradecer a mis tutores de TFM, Edu y Sonia, por su apoyo pero especialmente a Sonia, por su ayuda y completa disponibilidad durante el desarrollo de este trabajo, incluso cuando más ocupada estaba. Por haberme valorado tanto incluso cuando yo no lo hacía y por abrirme las puertas a la bioinformática. Es un lujo contar con profesoras tan apasionadas como tú. De corazón, gracias.

Resumen

Actualmente, el melanoma es el principal causante de muerte por cáncer de la piel. No todos los melanomas se comportan igual, por lo que es de interés disponer de una clasificación de los mismos que permita precisar mejor el pronóstico de la enfermedad y diseñar tratamientos más personalizados. Tradicionalmente, se ha clasificado a los pacientes de melanoma según el patrón de crecimiento y localización del tumor. Sin embargo, a día de hoy, gracias a la contribución de los investigadores en este campo, dicho enfoque ha cambiado y se tienen en cuenta distintas variables clínicas, epidemiológicas y genéticas que permiten caracterizar mejor la enfermedad, esto es, en conjunto, tienen en cuenta la etiopatogenia del melanoma del paciente.

La actual clasificación de la *Organización Mundial de la Salud* (OMS) permite catalogar a pacientes de melanoma en grupos etiopatogénicos. No obstante, en ciertos casos, la clasificación de los pacientes en dichos grupos etiopatogénicos no es posible a causa, por ejemplo, de no disponer de datos en variables decisivas para la clasificación o por tratarse de pacientes con patrones diferentes en las variables consideradas. Esto genera grupos etiopatogénicos que no están perfectamente definidos, por lo que es necesario caracterizarlos o redefinirlos mejor.

En el presente trabajo, se abordaron diferentes estrategias estadísticas para, por una parte, resolver la problemática de la falta de información en algunas variables y, por otra, redefinir mejor la clasificación para aquellos pacientes en grupos etiopatogénicos mal definidos o caracterizados. Para ello, se disponía de una base de datos proporcionada por el *Instituto Valenciano de Oncología* (IVO), con más de 2000 pacientes de melanoma en la que se incluyen una gran cantidad de datos clínicos, epidemiológicos y patológicos. Tras realizar un minucioso preprocesamiento y limpieza de los datos, se aplicaron técnicas de imputación de datos faltantes basadas en modelos predictivos que evitasen la máxima pérdida de información posible y se analizó y trató la existencia de valores anómalos que pudieran afectar a los análisis realizados. Seguidamente, se compararon diversas técnicas de *clustering* en busca de aquella que mejor clasificase los pacientes de etiopatogenia difusa. Entre las técnicas aplicadas, se encuentran los métodos jerárquicos, los de partición o incluso

el *clustering* difuso. Para finalizar, se realizó un estudio de caracterización de los *clusters* definidos, tanto desde el punto de vista multivariante (mediante PLSDA) como bivariante (tests de independencia) para relacionar los grupos etiopatogénicos con las variables utilizadas para el *clustering* y también para otras variables de interés como mutaciones en ciertos genes vinculados al melanoma. También se caracterizó la supervivencia en cada grupo etiopatogénico, mediante curvas Kaplan-Meier y modelos de regresión de Cox.

Abstract

Melanoma is currently the leading cause of death from skin cancer. Not all melanomas behave in the same way, so it is interesting to have a classification of melanomas that allows us to better determine the prognosis of the disease and to design more personalized treatments. Traditionally, melanoma patients have been classified according to growth pattern and location of the tumor. However, nowadays, thanks to the contribution of researchers in this field, this approach has changed and different clinical, epidemiological and genetic variables are taken into account to better characterize the disease, i.e., as a whole, they take into account the etiopathogenesis of the patient's melanoma.

The current *World Health Organization* (WHO) classification allows the assignment of melanoma patients into etiopathogenic groups. However, in some cases, the classification of patients into these etiopathogenic groups is not possible due to, for example, a lack of data on decisive variables for classification or due to the presence of different patterns in the variables considered, which results in etiopathogenic groups that are not perfectly defined, making it necessary to better characterize or redefine them.

In the current study, different statistical strategies were used to, on the one hand, solve the problem of the lack of information for some variables and, on the other hand, to redefine the classification of those patients in poorly defined or characterized etiopathogenic groups. To achieve this, a database provided by the *Instituto Valenciano de Oncología* (IVO), with more than 2000 melanoma patients was available, including a large amount of clinical, epidemiological and pathological data. After a meticulous preprocessing and cleaning of the data, missing data imputation techniques based on predictive models were applied to avoid the loss of information, and the existence of outliers that could affect the analyses performed was analyzed and treated. Afterwards, various techniques were compared seeking the one that best classified patients of diffuse etiopathogenesis. Among the techniques applied, we found hierarchical methods, partitioning methods or even fuzzy clustering. Finally, a characterization study of the defined clusters was carried out, both from a multivariate (using PLSDA) and bivariate (tests of independence) point of view to relate the etiopathogenic groups with the variables used for clustering and also for other

variables of interest such as mutations in certain genes linked to melanoma. Survival in each etiopathogenic group was also characterized using Kaplan-Meier curves and Cox regression models.

Resum

Actualment, el melanoma és el principal causant de mort per càncer de la pell. No tots els melanomes es comporten igual, per la qual cosa és d'interés disposar d'una classificació dels mateixos que permeta precisar millor el pronòstic de la malaltia i dissenyar tractaments més personalitzats. Tradicionalment, s'ha classificat als pacients de melanoma segons el patró de creixement i localització del tumor. No obstant això, hui dia, gràcies a la contribució dels investigadors en aquest camp, aquest enfocament ha canviat i es tenen en compte diferents variables clíniques, epidemiològiques i genètiques que caracteritzen millor la malaltia, això és, en conjunt, tenen en compte l'etiopatogènia del melanoma del pacient.

L'actual classificació de la *Organització Mundial de la Salut*(OMS) permet catalogar a pacients de melanoma en grups etiopatogènics. No obstant això, en certs casos, la classificació dels pacients en aquests grups etiopatogènics no és trivial a causa, per exemple, de no disposar de dades en variables decisives per a la classificació o per tractar-se de pacients amb patrons diferents en les variables considerades. Això genera grups etiopatogènics que no estan perfectament definits, per la qual cosa és necessari caracteritzar-los o redefinir-los millor.

En el present treball, es van abordar diferents estratègies estadístiques per a, d'una banda, resoldre la problemàtica de la falta d'informació en algunes variables i, per una altra, redefinir millor la classificació per a aquells pacients en grups etiopatogènics mal definits o caracteritzats. Per a això, es disposava d'una base de dades proporcionada pel el *Institut Valencià d'Oncologia* (IVO), amb més de 2000 pacients de melanoma en la qual s'inclouen una gran quantitat de dades clíniques, epidemiològics i patològics. Després de realitzar un minuciós preprocessament i neteja de les dades, es van aplicar tècniques d'imputació de dades perdudes basades en models predictius que evitaren la màxima pèrdua d'informació possible i es va analitzar i va tractar l'existència de valors anòmals que pogueren afectar les analisis realitzades. Seguidament, es van comparar diverses tècniques de *clustering* a la recerca d'aquella que millor classificara els pacients d'etiopatogènia difusa. Entre les tècniques aplicades, es troben els mètodes jeràrquics, els de partició o fins

i tot el *clustering* difús. Per a finalitzar, es va realitzar un estudi de caracterització dels *clusters* definits, tant des del punt de vista multivariant (mitjançant PLSDA) com a bivariate (tests d'independència) per a relacionar els grups etiopatogènics amb les variables utilitzades per al *clustering* i també per a altres variables d'interés com ara mutacions en alguns gens vinculats al melanoma. També es va caracteritzar la supervivència en cada grup etiopatogènic, mitjançant corbes Kaplan-Meier i models de regressió de Cox.

Nomenclatura

AEDV	Asociación Española de Dermatología y Venerología
ARI	<i>Adjusted Random Index</i>
CBC	Cáncer Basocelular Cutáneo
CEC	Cáncer Epidermoide Cutáneo
Cluster	Grupo
Complete linkage	Amalgamiento completo
CSD	<i>Cumulative Sun Damage</i>
FKM	<i>Fuzzy K-means</i>
Heatmap	Mapa de calor
HR	<i>Hazard Ratio</i>
IMC	Índice de Masa Corporal
IVO	Instituto Valenciano Oncológico
K-means	K-medias
K-medoids	K-medoides
MC1R	Gen receptor de melacortina-1
MICE	<i>Multivariate Imputation by Chained Equations</i>
n.a.	No aplicable
OMS	Organización Mundial de la Salud
PAM	<i>Partitioning Around Medoids</i>
PCA	<i>Principal Component Analysis</i>
PLSDA	<i>Partial Least Square Discriminant Analysis</i>
RHC	<i>Red Hair Color phenotype</i>
SCR	Suma de Cuadrados Residual
SCT	Suma de Cuadrados Total
Single linkage	Amalgamiento simple
TD	<i>Total Deviation</i>
TIL	<i>Tumor Infiltrating Lymphocytes</i>
TSD	<i>Total Sum of Distances</i>
v.f.	Valor faltante
VIP	<i>Variable Influence on Projection</i>
wt	<i>Wild type</i>

Índice general

Resumen	I
Abstract	II
Resum	III
Nomenclatura	IV
Índice de figuras	VIII
Índice de tablas	X
1. Introducción	1
1.1. Melanoma	1
1.2. Grupos etiopatogénicos	4
1.3. Epidemiología	6
1.3.1. Factores de riesgo	6
2. Objetivos	8
3. Material y Métodos	9
3.1. Descripción de la base de datos	9
3.2. Métodos	10
3.2.1. Pre-procesado de los datos	10
3.2.1.1. Limpieza	11
3.2.1.2. Imputación de datos faltantes	11
3.2.2. Exploración de datos mediante PCA	14
3.2.2.1. Validación del modelo PCA	15
3.2.3. Clustering	18
3.2.3.1. Medidas de distancia	19

ÍNDICE GENERAL

3.2.3.2. Tendencia de agrupamiento	22
3.2.3.3. Métodos de clustering	23
3.2.3.4. Determinación del número óptimo de clusters	26
3.2.3.5. Medidas de similitud entre clasificaciones	29
3.2.4. Caracterización de los clusters obtenidos: PLSDA	30
3.2.4.1. Validación del modelo PLSDA	32
3.2.4.2. Selección de variables del modelo PLSDA	32
3.2.4.3. Caracterización de los clusters	33
3.2.5. Análisis de supervivencia	34
3.3. Software	39
3.3.1. Librerías de R utilizadas	39
3.3.2. Desarrollo de funciones necesarias	40
4. Resultados	43
4.1. Limpieza de datos	43
4.2. Validación de la imputaciones mediante PCA	47
4.3. Análisis de valores anómalos mediante PCA	48
4.4. Clustering	51
4.4.1. Elección de la distancia y método de clustering	51
4.4.2. Obtención de clusters para pacientes de grupos 'indefinidos'	54
4.4.3. Caracterización de los clusters obtenidos	57
4.5. Relación de los grupos etiopatogénicos con variables de interés	60
4.5.1. Características histológicas pronósticas	60
4.5.2. Mutaciones somáticas	61
4.6. Relación de los clusters con la supervivencia	63
4.6.1. Curvas de Kaplan-Meier	63
4.6.2. Modelos de Cox para supervivencia global	64
4.6.3. Modelos de Cox para supervivencia específica	67
5. Conclusiones	69
Anexos	
A. Material y Métodos	74
A.1. Descripción de la base de datos	74
A.2. Marco teórico PCA	78

B. Resultados	81
B.1. Validación de las imputaciones	81
B.2. Clusters	83
B.2.1. Elección de número de clusters	83
B.2.2. Clusters obtenidos para los grupos etiopatogénicos bien definidos	85
B.3. Obtención de clusters para grupos 'indefinidos'	86
B.3.1. Comprobación resultados	88
B.4. Supervivencia	91
B.4.1. Estudios univariados	91
B.4.2. Estudio multivariado	94
C. Funciones desarrolladas	97
C.1. Limpieza base de datos	97
C.1.1. <i>dummytovariable</i>	97
C.2. Validación de las imputaciones	98
C.2.1. <i>plotloading</i>	98
C.2.2. <i>R2varcomp</i>	99
C.2.3. <i>SCR</i>	99
C.2.4. <i>T2</i>	100
C.2.5. <i>Contri</i>	100
C.3. Clustering	101
C.3.1. <i>hopkins</i>	101
C.3.2. <i>TD</i>	102
C.3.3. <i>Gower_pam_silhouette</i>	102
C.3.4. <i>diversity</i>	102
C.3.5. <i>TSD</i>	103
C.3.6. <i>fuzzyalgo</i>	103
C.4. PLSDA	104
C.4.1. <i>p.coef</i>	104
C.4.2. <i>plotweight</i>	105
D. Relación con los Objetivos de Desarrollo Sostenible	106
Bibliografía	107

Índice de figuras

1.1. Comparación de lunares normales y lunares asociados a melanoma.	2
1.2. Estimación de tasas de incidencia estandarizadas por edad en 2020, melanoma cutáneo, ambos sexos, todas las edades.	3
4.1. Gráfica de co-ocurrencia de v.f.	44
4.2. Distribuciones para las variables imputadas para el tipo de imputación 1.	45
4.3. Distribuciones para las variables imputadas para el tipo de imputación 2.	46
4.4. Resumen PCA para las bases de datos.	48
4.5. Gráficos SCR y T^2 -Hotelling para PCA sobre datos imputados.	49
4.6. Gráficos de <i>Scores</i> y <i>Loadings</i> de las componentes 1 y 3	50
4.7. Contribuciones a la <i>SCR</i> del individuo 1139	50
4.8. <i>Heatmaps</i> para los pacientes de grupos etiopatogénicos bien definidos según las distancias.	51
4.9. <i>Heatmap</i> con la distancia de Manhattan para pacientes de grupos 'indefinidos'.	54
4.10. Proyecciones de los pacientes de grupos 'indefinidos' sobre las dos primeras componentes principales para métodos escogidos.	55
4.11. Boxplots para la tasa de error de clasificación balanceada para los distintos números de componentes, obtenido por <i>cross validation</i> , para las opciones con 3 <i>clusters</i> y 3 <i>clusters</i> más un grupo difuso.	57
4.12. <i>Weightings</i> de las componentes 1 y 2 para el modelo PLSDA creado. . . .	58
4.13. Curvas de Kaplan-Meier para el análisis de supervivencia para los grupos etiopatogénicos bien definidos y los grupos encontrados.	63
4.14. Residuos <i>deviance</i> para el modelo multivariante de supervivencia global creado a partir de regresión de Cox.	65
4.15. Beta dependiente del tiempo para la variable <i>Mitosis</i>	65

B.1.	<i>Loading plots</i> y R^2 explicada por cada variable en cada componente para las distintas bases de datos.	82
B.2.	Métodos de selección de número de <i>clusters</i> para distancia de Manhattan método <i>K-means complete</i>	83
B.3.	Métodos de selección de número de <i>clusters</i> para distancia de Gower método <i>K-means Ward</i>	84
B.4.	Métodos de selección de número de <i>clusters</i> óptimo para distancia de Gower método <i>K-medoids</i>	84
B.5.	Métodos de selección de número de clusters óptimo para método <i>K-prototypes</i>	84
B.6.	Métodos de selección de número de <i>clusters</i> para la distancia de Manhattan mediante el método <i>K-means</i> para pacientes de grupos etiopatogénicos difusos.	87
B.7.	Coeficiente de Silhouette para las diferentes técnicas para pacientes de grupos 'indefinidos'.	87
B.8.	Curvas de Kaplan-Meier para estudio de supervivencia según sexo.	91
B.9.	Curvas de Kaplan Meier para estudio de supervivencia por regresión.	91
B.10.	Curvas de Kaplan-Meier para estudio de supervivencia para satelitosis.	92
B.11.	Curvas de Kaplan-Meier para estudio de supervivencia para ulceración.	92
B.12.	Curvas de Kaplan Meier para estudio de supervivencia según TIL.	93
B.13.	Curvas de Kaplan-Meier para estudio de supervivencia para Invasión vascular.	93
B.14.	Curvas de Kaplan-Meier para estudio de supervivencia según Breslow.	94
B.15.	Curvas de Kaplan-Meier para estudio de supervivencia para Mitosis.	94
B.16.	Curvas de Kaplan-Meier para estudio de supervivencia según Total de ganglios positivos.	95
B.17.	Residuos deviance para el modelo multivariante de supervivencia global.	95
B.18.	Residuos deviance para el modelo multivariante de supervivencia específica.	96
B.19.	Residuos deviance para el modelo multivariante de supervivencia global para los grupos etiopatogénicos juntados	96

Índice de tablas

3.1. Verificaciones y transformaciones sobre la base de datos.	11
4.1. Resumen sobre porcentaje de v.f. para los individuos.	43
4.2. Resumen estadístico de Hopkins para distintas distancias para grupos etiopatogénicos bien definidos.	52
4.3. Resultados de las medidas de similitud para los diferentes métodos de clustering aplicado a individuos con grupos etiopatogénicos bien definidos.	53
4.4. Resumen estadístico de Hopkins para distancia de Manhattan para grupos 'indefinidos'.	55
4.5. Tabla de confusión real frente a predicho del modelo PLSDA creado para la clasificación de los pacientes de grupos 'indefinidos' en los 3 <i>clusters</i> . .	58
4.6. Análisis univariado de las variables externas a la creación de los <i>clusters</i> en relación con los grupos etiopatogénicos bien definidos y los grupos creados.	60
4.7. Análisis univariado de las mutaciones somáticas en relación con los grupos etiopatogénicos bien definidos y los grupos creados.	62
4.8. Residuos <i>Schoenfeld</i> para el modelo multivariante de supervivencia global creado a partir de regresión de Cox.	65
4.9. Modelo de supervivencia global.	66
4.10. Modelo de supervivencia global juntando los grupos de perfiles etiopatogénicos parecidos.	67
4.11. Modelo de supervivencia específica.	68
B.1. Distribución del grupo etiopatogénico en los <i>clusters</i> para el método <i>K-means</i> mediante distancia de Manhattan.	85
B.2. Distribución del grupo etiopatogénico en los <i>clusters</i> para el método de <i>Fuzzy K-means</i>	85

B.3. Distribución del grupo etiopatogénico en los <i>clusters</i> para el método jerárquico mediante distancia de Manhattan	86
B.4. Pacientes según <i>Grupo etiopatogénico</i> y grupos creados	88
B.5. Distribuciones de las variables en los grupos creados	89
B.6. Distribuciones de las variables en los grupos creados. Continuación Tabla B.5	90
B.7. Residuos <i>Schoenfeld</i> para el modelo multivariante de supervivencia global.	95
B.8. Residuos <i>Schoenfeld</i> para el modelo multivariante de supervivencia específica.	96
B.9. Residuos <i>Schoenfeld</i> para el modelo multivariante de supervivencia global para los grupos etiopatogénicos juntados	96

CAPÍTULO 1

Introducción

1.1. Melanoma

El cáncer es una de las principales causas de muerte a nivel global. Según informa la Organización Mundial de la Salud (OMS), en 2020 se convirtió en la segunda causa de muerte cobrándose la vida de 10 millones de personas y diagnosticando más de 19 millones de casos nuevos. Aunque en la mayoría de los casos esta enfermedad tenga tratamiento, es imprescindible el diagnóstico precoz de la enfermedad y su tratamiento. Pese a ello, y según un reciente sondeo de la OMS, a causa de la pandemia de la COVID-19 los tratamientos de esta enfermedad se interrumpieron en más del 40 % de los países.

Según datos de la Agencia Internacional para la investigación del Cáncer (IARC), se estima que para 2040 la incidencia del cáncer se vea incrementada en un 49 %. Sin embargo, la incidencia esperada no es la misma para todos los tipos de cáncer. En el presente trabajo se tratará uno de los tipos de cáncer que se estima tendrá una de las mayores subidas en la incidencia, el melanoma cutáneo. Pese a ser el menos común entre los tipos de cáncer de piel (en 2016 se estimó que constituía tan solo el 4 % de entre los canceres dermatológicos), el melanoma es el causante de la mayoría de muertes debidas al cáncer de piel (en el mismo periodo se estimó que causó el 80 % de las muertes debidas a cáncer de piel).

De acuerdo con el secretario general de la Asociación Española de Dermatología y Venereología (AEDV), Luis Ríos, el melanoma se convertirá en 2040 en el segundo tumor con mayor incidencia global, y el primero entre los hombres por delante del cáncer de colon y el de pulmón [Galván, 2022].

En 1806, el médico francés René Laennec describió el melanoma por primera vez como una enfermedad y fue en 1812 cuando publicó un artículo en el Boletín de la Facultad de Medicina de París donde utilizó el término *Mélanose* para referirse a él, palabra que deriva del griego 'negro'.

El melanoma [Casanova Seuma JM, 2004] es un tumor maligno derivado de los melanocitos, células que producen la melanina, el pigmento marrón de la piel responsable del color de la piel y los ojos, además de proteger las capas más profundas de la piel contra los efectos nocivos de la radiación ultravioleta. Para ser más precisos, el melanoma es un tipo de cáncer de la piel que se desarrolla cuando los melanocitos empiezan a crecer fuera de control.

Aunque sea el menos común entre los cánceres de piel, el melanoma es más peligroso que cualquier otro puesto que es mucho más probable que se extienda a otras partes del cuerpo si no se detecta y trata a tiempo.

Los melanomas pueden desarrollarse en la piel o la mucosa de cualquier parte del cuerpo, aunque los lugares más comunes son la espalda entre los hombres y las piernas entre las mujeres. A su vez, suelen ser poco comunes en áreas protegidas a la exposición solar como las mucosas oral, ocular o anal.

En la mayoría de los casos, los melanomas pueden identificarse clínicamente por tratarse de lunares asimétricos, con bordes irregulares, tener varios colores, ser más grandes de lo normal, sangrar de vez en cuando o por tratarse de lunares que cambian gradualmente de forma, tamaño o color.



Figura 1.1: Comparación de lunares normales y lunares asociados a melanoma.

Los autoexámenes de la piel en busca de anomalías como las presentadas en la Figura 1.1 pueden ser útiles para la detección temprana del melanoma. De hecho, es más fácil curar los tumores y más improbable que causen la muerte cuando estos no han crecido se encuentran en un estadio temprano.

Con el fin de facilitar estos autoexámenes, se han desarrollado ciertas apps como *eDermma*, *SkinVision* o *AI Dermatologist* que con la ayuda de la inteligencia artificial permiten identificar los lunares que deberían ser evaluados por un especialista con tal de detectar, por ejemplo, posibles lunares relacionados al melanoma.

Además de las diferencias entre las localizaciones de aparición y la morfología de los tumores, como se recoge en la Figura 1.2, la incidencia del melanoma cutáneo no es la misma en todos los países, siendo más prevalente en aquellos de población de raza caucásica. Ejemplo de ello son países como Australia o Nueva Zelanda donde se encuentra la mayor incidencia de melanoma y cuya población vive en un área con gran índice de radiación solar. Por tanto, las tasas de incidencia también sugieren que la radiación solar juega un papel fundamental en el desarrollo del melanoma.

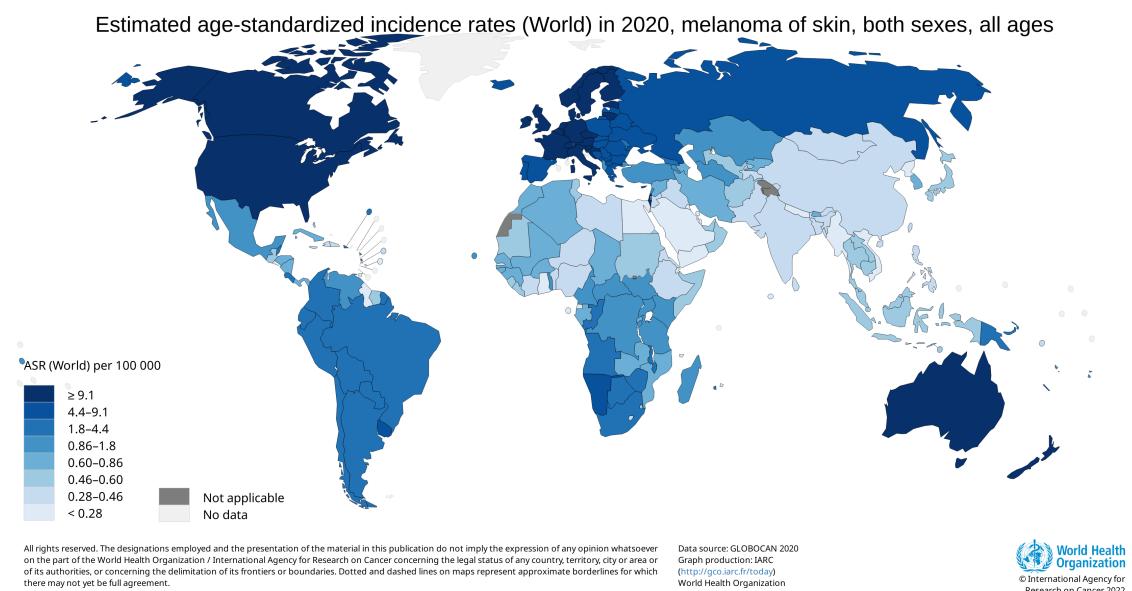


Figura 1.2: Estimación de tasas de incidencia estandarizadas por edad en 2020, melanoma cutáneo, ambos sexos, todas las edades.

Fuente: *Global Cancer Observatory* (GLOBOCAN), 2020.

En vista de los datos, son muchos los trabajos que se han realizado en busca de variables clínicas, epidemiológicas, histopatológicas y de biología molecular que ayuden a comprender la incidencia del melanoma.

1.2. Grupos etiopatogénicos

La etiopatogenia hace referencia al origen o causa del desarrollo de una enfermedad. Con los años, el estudio de la etiopatogenia se ha vuelto de vital importancia para entender mejor quiénes tienen mayor probabilidad de padecer ciertas afecciones. Para ello, entre otros, estudia los hábitos del paciente y el entorno que le rodea. Y es que este estudio es de interés dado que como se puede analizar en la Figura 1.2 no todos los seres humanos somos igual de propensos a padecer melanoma.

La creación y búsqueda de grupos etiopatogénicos podría por tanto ayudar a los médicos a recomendar exámenes de detección precoz a cierto tipo de pacientes y el tipo de pruebas y frecuencia en la que realizarlas para combatirlo. No obstante, la definición de estos grupos etiopatogénicos por criterios médicos o mediante técnicas estadísticas puede suponer un gran desafío por la naturaleza de los datos médicos. Los datos suelen ser recogidos en consulta en poco tiempo (como consecuencia pueden contener errores), por diferentes especialistas que pueden variar en criterios o nomenclaturas, suelen estar altamente correlacionados y contienen un alto porcentaje de valores perdidos que no se recogen debido a la condición médica, la pérdida de muestras o incluso la falta de tiempo.

El grupo de investigación en melanoma del IVO, basados en los conceptos incluidos en las más recientes recomendaciones de la OMS y en trabajos previos donde se han tratado de identificar grupos etiopatogénicos para pacientes de melanoma, han propuesto una clasificación para estos pacientes según su etiopatogenia. En dicha clasificación se definen los siguientes grupos etiopatogénicos:

- **Primario desconocido.** Este grupo incluye a pacientes de los que se desconoce cuál fue la localización primaria (la parte del cuerpo en la que se originó) del melanoma. Esto ocurre cuando se encuentra cáncer en una o más localizaciones y no se puede determinar el lugar primario. Estrictamente, este grupo no tiene una etiopatogenia definida puesto que no puede tener en consideración algunas características del tumor primario.
- **CSD.** Lo constituyen pacientes que se caracterizan por tener el melanoma asociado con el daño solar acumulado (del inglés *cumulative sun damage*, CSD), expresado con la presencia de al menos grado moderado de Elastosis¹ en la piel donde se

¹Elastosis: Signos de envejecimiento o degradación en la dermis circundante al melanoma por la exposición solar acumulada

desarrolla el tumor. Sus melanomas se asocian a localizaciones expuestas a la radiación solar crónica como la cabeza o el cuello. Además, se ha encontrado cierta asociación con mutaciones en el gen KIT.

- **Acral.** Este grupo está compuesto por pacientes con melanoma en piel de tipo acral; es decir, en la planta del pie, en las palmas de las manos o en las uñas. Se caracteriza por aparecer en pacientes mayores, con pocos nevus atípicos y poca acumulación de daño solar. Son de subtipo histológico melanoma acral lentiginoso. Su pronóstico suele ser peor que el de otros grupos etiopatogénicos. Además, se ve poco relacionado con mutaciones del gen BRAF, pero tienen cierta prevalencia mutaciones del gen KIT.
- **Mucoso.** Se forma por pacientes con melanoma en las membranas mucosas de la cavidad oral, respiratoria, gastrointestinal y/o del tracto genitourinario. Son melanomas poco frecuentes y se caracterizan por tener un pronóstico peor que el melanoma cutáneo.
- **Nevogénico.** Este grupo incluye pacientes con una mayor predisposición intrínseca a la proliferación melanocítica. Como consecuencia tienen un número elevado de nevos melanocíticos (tumores benignos derivados de los melanocitos). Aunque en la literatura no existan criterios que de forma precisa indiquen qué número de nevus melanocíticos se debería considerar como expresión de esta predisposición, se consideró como nevogénico aquél paciente que tuviera un número total superior a 50. Este grupo suele estar formado por pacientes jóvenes y con melanoma en áreas en las que la exposición solar es intermitente.
- **Nevogénico débil.** La diferencia con el grupo etiopatogénico nevogénico es sutil. Esta diferencia se basa en el hecho de que no existe realmente un punto de corte para el número de nevus que un individuo debe tener para que se trate de melanoma nevogénico. Por tanto, es una clasificación creada por el experto que incluye a aquellos pacientes que tengan entre 20 y 50 nevus.
- **Mixto.** Pacientes que tienen perfil con características definitorias tanto de nevogénico como de CSD.

Sin embargo, no todos los pacientes son fácilmente clasificables en los 7 grupos etiopatogénicos que se acaban de introducir debido a la falta de valores en variables decisivas para la discriminación. Para los pacientes de los que la etiopatogenia resulta todavía difusa se

distingue entre aquellos de los que no se disponen todos los datos para su clasificación (los clasificados como 'No clasificable') y aquellos que no tienen ninguna de las características típicas de los grupos etiopatogénicos anteriormente definidos (los clasificados como *Non-risky*). Este último grupo de pacientes, es de gran relevancia por tratarse de individuos que desarrollan melanoma aunque no presenten ningún factor de riesgo conocido. Será por tanto de interés, la creación de unos grupos que consigan clasificar o reagrupar a los pacientes que hasta el momento se han considerado como 'No clasificable' o *Non-risky* según su etiopatogenia. Es en este interés en el que se basa el presente Trabajo de Fin de Máster.

1.3. Epidemiología

En las últimas décadas, la incidencia del melanoma no ha cesado de incrementar en la población de raza caucásica [Guy et al., 2015; UK, 2022; Liszkay et al., 2021]. Como se recogía en la Figura (1.2) Australia, Nueva Zelanda, Noruega, Suecia y Estados Unidos son algunos de los países que mayor incidencia presentaron de melanoma en 2020. España pese a presentar una de las tasas más bajas de incidencia y mortalidad de Europa, ha multiplicado por 10 su incidencia en los últimos 20 años [Sáenz et al., 2005]. En cuanto a la mortalidad se refiere, en las últimas décadas también se han registrado incrementos tanto en España como en el resto de países; sin embargo, los datos parecen mostrar cierta estabilización.

Con el fin de reducir la incidencia y mortalidad de esta enfermedad, es de vital importancia la investigación sobre los factores de riesgo, las medidas preventivas y los tratamientos.

1.3.1. Factores de riesgo

La exposición solar es una de las principales causas en el desarrollo del melanoma [Ródenas et al., 1996; Berwick et al., 2016]. La exposición excesiva al sol, principalmente en la infancia representa uno de los principales factores de riesgo de cáncer de piel [Cercato et al., 2013]. Según el estudio suele variar la manera en la que se considera la exposición siendo la clasificación más común la de intermitente (debida a actividades de ocio), crónica (como consecuencia del trabajo) o total (resultado de ambas prácticas).

La exposición solar suele en ciertos casos dar lugar a otro factor de riesgo, las quemaduras solares. Con diferentes efectos según la edad del paciente [Dennis et al., 2008].

Estudios demuestran que la exposición solar puede causar melanoma en todo el cuerpo; no obstante, se considera la localización del tumor factor de riesgo puesto que no todas las localizaciones presentan la misma incidencia, siendo aquellas expuestas al sol de forma intermitente las que mayor incidencia presentan [Nagore et al., 2009].

Otros factores de riesgo a tener en cuenta son aquellos relacionados con el fenotipo del paciente (como el color de pelo y ojos), el fototipo (capacidad de la piel de asimilar radiación solar), sexo o el recuento y tipo de nevus (lunares) del paciente [Berwick et al., 2016].

Aunque la exposición solar juegue un papel fundamental en el desarrollo del melanoma, cabe recordar que la tercera parte de los melanomas que se producen no parecen estar relacionados con la exposición solar [Sáenz et al., 2005]. Cada vez más estudios apuntan a que el efecto de la exposición solar no sea el mismo para todas las personas, por ello es importante tener en cuenta los factores de riesgo relacionados con la genética.

Son diversos los genes que se están estudiando por su posible relación con el desarrollo del melanoma. Entre otros, se han identificado ciertos genes relacionados con la pigmentación [Berwick et al., 2016] como el gen receptor de melanocortina-1 (MC1R) o alguna de sus variantes referidas como fenotipo pelirrojo (*Red Hair Color phenotype* en inglés o también conocido por sus siglas RHC).

Otras vías de investigación apuntan hacia el perfil molecular de los melanomas, en particular las mutaciones en los genes BRAF y NRAS, las más prevalentes, que reflejan diferentes patrones de asociación con diversos factores de riesgo del paciente [Hacker et al., 2012].

En el presente trabajo, será por tanto de vital importancia el estudio de los factores de riesgo presentados y su relación con los distintos grupos etiopatogénicos definidos.

CAPÍTULO 2

Objetivos

En este trabajo se establecieron dos objetivos principales y varios objetivos específicos a llevar a cabo para conseguirlos:

1. Creación de grupos etiopatogénicos según características clínicas, histológicas, epidemiológicas y mutacionales para pacientes de melanoma que no pueden clasificarse en uno de los siete grupos etiopatogénicos propuestos por el IVO.
 - Limpieza e imputación de la base de datos para evitar la pérdida de información.
 - Comparación y selección de métodos de clustering para la obtención de los grupos.
2. Caracterización de los grupos etiopatogénicos encontrados e identificación de las características clínicas, histológicas, epidemiológicas y mutacionales que los diferencian.
 - Aplicación de técnicas multivariadas para la identificación de variables decisivas en la creación de los grupos etiopatogénicos.
 - Analizar e identificar los factores protectores ante la supervivencia de los pacientes de melanoma.

CAPÍTULO 3

Material y Métodos

3.1. Descripción de la base de datos

La base de datos utilizada en el presente trabajo estaba compuesta por pacientes diagnosticados de melanoma de la Unidad de dermatología del IVO y fue creada manualmente en la propia consulta mientras que los pacientes eran examinados y contestaban preguntas que podrían ser relevantes. Dicha base de datos, está formada por 2304 individuos y 84 variables las cuales podrían reagruparse en 6 bloques de variables a tratar. A continuación se presentan las variables analizadas aunque para mayor información véase Anexo A.1.

- **Variables clínico-epidemiológicas.** Se utilizan para describir y caracterizar la población en riesgo de cierta enfermedad. En nuestro caso las variables tratadas fueron: Sexo, Edad, Índice de Masa Corporal (IMC), Fototipo, Ojos, Pelo, Quemaduras, Años de exposición a radiación solar debido a profesión, Tabaquismo, Efélides, Lentigos, Queratosis actínicas, Segundo tumor, Carcinoma Epidermoide Cutáneo (CEC), Carcinoma Basocelular (CBC), Angiomas seniles, Queratosis seborreicas, Nevus melanocíticos comunes, Nevus melanocíticos atípicos, Múltiples melanomas, Melanoma familiar, Cáncer familiar, MC1R, RHC y R163Q.
- **Variables clínico-patológicas de caracterización del melanoma.** Describen las características del melanoma del paciente, analizando los síntomas y signos a través de los cuales se manifiesta dicha enfermedad así como las causas que lo producen. En este caso las variables analizadas fueron: Quemaduras en área del melanoma, Lentigos en área del melanoma, Localización, Fotolocalización, Tipo histológico, Restos de nevus en el melanoma, CSD y Elastosis.

- **Mutaciones somáticas.** Genes característicos del melanoma. Se analizaron los genes: BRAF, NRAS, KIT y promotor del TERT.
- **Características histológicas pronósticas.** Contiene información sobre el tejido del melanoma. Las estudiadas en este caso fueron: Breslow, Ulceración, Linfocitos infiltrantes en el tumor (del inglés *Tumor Infiltrating Lymphocytes*, TIL), Satelitosis, Mitosis, Regresión, Invasión vascular, Ganglio centinela y Total de ganglios positivos.
- **Estadio.** Clasificación según el melanoma está localizado solo en la piel, donde se desarrolló o si ya ha desarrollado metástasis bien vía linfática, bien vía hematogena.
- **Datos para estudio de supervivencia.** Contienen información sobre fechas de diagnóstico del melanoma, estatus del paciente o fechas de pérdida de seguimiento.
- **Grupo etiopatogénico.** Variable de interés en el trabajo que se lleva a cabo. Clasificación creada por el médico encargado de la base de datos.

Dada la naturaleza de los datos y de su extracción, la base cuenta con numerosos datos faltantes, más o menos frecuentes según el tipo de variables.

3.2. Métodos

En este apartado, se introduce la metodología aplicada para el correcto desarrollo del trabajo de fin de máster y los aspectos teóricos necesarios para poder llevarla a cabo.

3.2.1. Pre-procesado de los datos

Como en cualquier proyecto de análisis de datos, antes de empezar a analizar los datos se comenzó realizando un pre-procesado de los mismos. Este punto cobró gran relevancia puesto que, como en muchos otros contextos, la base de datos contaba con gran cantidad de valores faltantes (v.f.) y de valores que podían ser erróneos, que solo podrían detectarse gracias al minucioso tratamiento de los datos y la comunicación estrecha con el experto.

El pre-procesamiento de datos es el conjunto de todas aquellas técnicas que transforman los datos crudos en útiles. Su correcta ejecución condicionará los resultados extraídos de los estudios que le suceden, de ahí radica la importancia de un correcto pre-procesamiento y por ello supone más del 50 % del trabajo en muchos estudios, entre los que se incluye este.

3.2.1.1. Limpieza

Las bases de datos muchas veces suelen contener v.f., erratas, datos irrelevantes y/o discrepancias. Detectar estos problemas y corregir los datos crudos, es uno de los retos en el análisis de datos. De no hacerlo correctamente, podría resultar en un análisis incorrecto y consecuentemente una toma de decisiones poco fiable. Es por ello que la limpieza de datos es la etapa más importante dentro del pre-procesado de los datos.

Para llevar a cabo dicho tratamiento de datos se comenzó por realizar ciertas verificaciones y transformaciones sobre la base de datos que se recogen en la Tabla 3.1. Cabe aclarar que en el presente trabajo se denotan por n.a. a los valores 'No Aplicable'.

Tabla 3.1: Verificaciones y transformaciones sobre la base de datos.

	Variable	Apuntes	Solución
Verificaciones	ID	Identificadores no duplicados	Corregido por el experto
	Fototipo	No debe haber albinos (0)	v.f.
	Añospaquete	Añospaquete >130 imposible	v.f.
	LentigosareaMM	No Lentigos ⇒ No LentigosareaMM	Corregido por el experto
	Quemaduras	No Quemaduras ⇒ No QuemadurasareaMM graves	Corregido por el experto
	RHC	RHC ≤ MC1R	Corregido por el experto
	Grupo etiopatogénico	No hay OCULAR	Eliminarlos
	Estadio	Sateliosis GanglioCentinela=2⇒Estadio=2	
Transformaciones	Estadio	Breslow ≠ 999 ⇒ Estadio ≠ 0	
	Breslow		888 y 999→v.f.
	Cáncer\Páncreas\Melanoma Familiar		88 y 99→ v.f.
	Añosprofsol\Añospaquete\Peso		999→v.f.
	Resto de variables		99 y 999→ v.f.
	Añosprofsol\Añospaquete\Peso		777 y 888 → 0
	Breslow\Mitosis\QuemadurasareaMM		777 → n.a. (-1)
	Breslow		Breslow v.f. y Estadio 0 → Breslow in situ (-2)
	TIL		777→ n.a. (-2)
	Resto de variables		77,88,777 y 888 → n.a. (88)
	Cáncer Familiar	Eliminar las variables auxiliares	Cáncer\Páncreas Familiar no → Cáncer Familiar no
	IMC	Eliminar las variables auxiliares	$IMC = Peso/Altura^2$

3.2.1.2. Imputación de datos faltantes

En este punto era importante detectar, si los hubiera la presencia de **datos faltantes**.

Al manejar datos médicos, es frecuente la ausencia de datos debido a que pueden ser datos recogidos en consultas de manera fugaz o por contar con tiempo limitado entre consultas que puede dar pie a la falta de recogida de datos necesarios. Estas son algunas de las razones por las que surgen los datos faltantes. Hay dos enfoques a la hora de manejar la falta de datos en una base de datos: eliminarlos o imputarlos.

El primer caso solo es viable cuando la cantidad de datos faltantes no es excesiva y se posee una base de datos lo suficientemente grande. Si las condiciones necesarias se cumpliesen, sería un buen enfoque dado que los datos que se manejan no están expuestos a

posibles variaciones. En este caso, la existencia de algún dato faltante en alguna de las variables llevaría a la eliminación de toda la observación. En el caso adverso, cuando no se cumplan las condiciones necesarias, la eliminación de observaciones puede llevar a la pérdida de gran parte de la información, a la falta de potencia estadística en los análisis y a la posterior extracción de conclusiones erróneas. En estos casos, la imputación de datos faltantes suele ser más fiable aunque esté sujeta a una pequeña variación en los datos.

Una vez efectuadas las verificaciones y transformaciones comentadas en la Tabla 3.1, y dejando por el momento de lado las variables relacionadas con las mutaciones somáticas y las de supervivencia, se procedió a la imputación de datos para evitar la pérdida de pacientes.

Pese a que haya gran cantidad de técnicas para la imputación de datos (como la extendida imputación mediante la media), en esta sección tan solo se incluirá la imputación por regresión, la utilizada en este trabajo.

Imputación por regresión

Esta técnica reemplaza cada uno de los valores perdidos de una variable con una estimación obtenida tras crear un modelo de regresión del resto de variables de la base de datos. El modelo de regresión suele estimarse a partir de los datos observados y seguidamente el modelo suele usarse para la imputación de los v.f. Al tratarse de un modelo creado por regresión, hay que tener especial cuidado con variables correlacionadas para que el modelo no sea inestable y produzca estimaciones precisas.

Seleccionada la técnica de imputación a aplicar en la base de datos, antes de comenzar con la imputación y para impedir la errónea aplicación de la técnica, se realizó un análisis sobre los v.f. tanto de los individuos como de las variables. Del estudio se eliminó todo individuo que tuviese más de un 20 % de v.f., además se excluyó de la imputación toda variable que contuviese más de un 20 % de v.f.. Se optó por dicho porcentaje de v.f. como límite tras haber analizado la cantidad de pacientes que se eliminarían para distintos límites y tomar una decisión conjunta con el experto. Aunque ciertas variables pudiesen ser excluidas de la imputación, no se contempló su eliminación del estudio, por tanto si hubiese variables que superasen el límite impuesto, y para que estos no condujesen a pérdida de información en los análisis posteriores, dichos valores se tratarían como desconocidos. En el caso de las variables categóricas, se crearía un nuevo nivel para los desconocidos. En cambio, en el caso de las variables numéricas, al desconocido se le asignaría un valor fuera del rango de la variable original pero que fuese cercano a este para que no modificase la distribución original. Cabe aclarar que el estudio de v.f. tanto

en individuos como en variables se llevó a cabo de forma simultanea. Así, en caso de que alguna variable presentase una basta cantidad de v.f. se excluiría del estudio a realizar sobre los individuos. De esta manera, no se descartarían individuos de forma innecesaria ya que las variables con alto porcentaje de v.f. no podrían ser imputadas y por tanto sus v.f. no contribuirían al porcentaje de v.f. del individuo.

Finalmente, se inició la imputación del resto de variables no sin antes haber realizado un previo análisis de co-ocurrencia de v.f. y tener en cuenta la posible correlación entre variables para garantizar la idoneidad de las imputaciones (dado que, como en cualquier modelo de regresión, variables correlacionadas pueden dar lugar a estimaciones inestables).

En este estudio se optó por realizar la imputación mediante la librería *mice*. La librería imputa los v.f. mediante modelos de regresión y permite obtener m imputaciones por cada vez que se aplica. Haciendo uso de esta funcionalidad, se le exigieron 5 imputaciones para la posterior elección de aquella que presentase distribuciones más parejas a los datos de partida. Para cada variable a imputar *mice* creará un modelo propio en el que será importante tener en cuenta la co-ocurrencia de v.f. y la correlación de las variables analizadas anteriormente. Para ello, primero se aplicó la función *quickpred* de *mice* que permite crear una matriz de predictores usando el proceso de selección de variables descrito en [van Buuren et al., 1999] y que ayudará a establecer un buen modelo de imputación para datos con muchas variables. Aunque se pre-calculase la matriz de predictores, atendiendo a las correlaciones observadas y las relaciones conocidas entre variables que indicó el experto, se manipuló dicha matriz de predictores para que esta no contuviese variables correlacionadas. Así mismo, se empleó la imputación pasiva para la variable IMC, indicando al programa que dicha variable se construía a partir de las variables Peso y Altura y que debía excluirse a la hora de imputar las variables Altura y Peso para evitar la circularidad. También se realizó un post-procesado para limitar las imputaciones a valores posibles; es decir, se evitaron las imputaciones de valores 'No Aplicable' (n.a.) y se obligó a que cumpliesen las verificaciones de la Tabla 3.1.

Evitar la imputación de n.a. supuso un gran reto que se abordó tomando dos vías diferentes para posteriormente elegir la mejor de ellas. En la primera, se optó por reemplazar los n.a. con el valor más frecuente. La segunda opción fue un tanto más compleja. Primero se crearon $k-1$ *dummies* para las variables de k categorías (una de ellas correspondiente a tener o no n.a.). Una vez creadas, se aplicó la función *quickpred* en busca de la matriz de predictores y se eliminaron una vez más las variables correlacionadas en cada caso. No obstante, en este caso, también se eliminaron como predictores las *dummies* correspon-

dientes a los n.a. Al no existir como predictores, para el modelo nunca existieron valores n.a. por tanto en ningún caso los imputaría. Terminada la imputación, se juntaron las *dummies* en una única variable como lo estaban originalmente (esto fue posible mediante la función *dummytovariable* incluida en el Anexo C). Cuando se finalizó con la imputación, en ambos casos, se estudiaron las distribuciones de las variables (mediante las gráficas *stripplot* y *propplot* que aporta *mice*) con tal de elegir aquella que podría asemejarse más a la distribución de partida.

3.2.2. Exploración de datos mediante PCA

Terminada la imputación mediante las dos opciones comentadas en la sección anterior, se exploraron los datos imputados mediante PCA para comprobar la idoneidad de las mismas.

El análisis de componentes principales (PCA) [Abdi and Williams, 2010; Kherif and Latypova, 2020] es una técnica multivariada no supervisada que analiza un conjunto de datos en el que las observaciones están descritas por un conjunto de variables, en la mayoría de casos correlacionadas, cuantitativas. El PCA pertenece a la familia de los métodos de reducción de dimensionalidad y es frecuentemente utilizado en el tratamiento de bases de datos muy grandes y altamente correlacionadas. Otras aplicaciones en las que se suelen usar esta técnica incluyen, la eliminación de ruido en las señales, la separación ciega de fuentes y la compresión de los datos.

Esta técnica tiene como objetivo extraer la información relevante y expresarla en un subespacio de menor dimensionalidad, mediante un nuevo conjunto de variables ortogonales llamadas 'Componentes principales', que se obtienen mediante combinaciones lineales de las variables originales.

Para extraer cada una de esas componentes, los datos originales se proyectan en un nuevo sistema de coordenadas, donde se buscan los ejes ortogonales ('ejes principales') a lo largo de los cuales los datos varíen lo máximo posible; siendo el primer eje en el que mayor variabilidad se exprese, el segundo el que mayor variabilidad refleje tras el primer eje y así sucesivamente hasta que la matriz de datos original se descomponga. Las componentes principales son las proyecciones de los datos originales sobre los ejes principales extraídos de la manera anteriormente explicada, expresándose por tanto la mayor varianza de los datos en la primera componente principal. Las componentes principales se obtienen deconstruyendo la matriz de datos en vectores propios (los ejes principales) y sus corres-

pondientes valores propios (la varianza explicada por el eje principal correspondiente). Para la explicación matemática e introducción de datos discretos en esta técnica véase Anexo A.2.

Para analizar la idoneidad de las imputaciones, se realizó un PCA sobre cada una de las bases de datos imputadas y sus resultados fueron comparados con los PCA obtenidos sobre la base de datos original sin imputar, primero excluyendo pacientes con v.f. y después utilizando el método NIPALS, que permite utilizar bases de datos incompletas. Cabe aclarar que en los modelos PCA tan solo se incluyeron las variables que fueran a utilizarse posteriormente en el *clustering*.

La comprobación a realizar, comenzó por analizar si las imputaciones daban lugar a alguna observación extrema y/o atípica no propia de la naturaleza de las observaciones. Una vez analizado esto, se exploraron los *loadings* correspondientes (mediante la función *plotloading* incluida en el Anexo C) ayudándose también de los gráficos de la R^2 explicada por cada variable en cada componente (creados con la función *R2varcomp* e incluida en el mismo Anexo). Mediante los *loadings*, se analizó si las relaciones encontradas en la base de datos original se mantuvieron una vez realizadas las imputaciones. Analizados todos los aspectos, se optaría por aquella imputación que no diese lugar a observaciones anómalas consecuencia de la imputación y por aquella que preservase las relaciones de la base de datos original. En caso de que ninguna de ellas presentase diferencias respecto a la otra, habría que aplicar algún otro criterio para elegir entre ellas.

Para la completa comprensión de las técnicas anteriormente mencionadas véase la sección inmediata. Dichas técnicas no han sido previamente definidas por tratarse de técnicas propias de la validación de los modelos PCA aunque en este caso no se hayan utilizado con dicho fin.

3.2.2.1. Validación del modelo PCA

Antes de entrar en mayor detalle sobre la validación de los modelos PCA, caben realizar ciertos apuntes teóricos para la completa comprensión del lector de las técnicas utilizadas.

En el Anexo A.2 se introduce cómo se puede proyectar la base de datos mediante la ecuación (A.1), es trivial por tanto que $\mathbf{X} = \mathbf{T}\mathbf{P}^T$. Sin embargo, esta igualdad es tan solo cierta cuando se extraen todas las componentes principales posibles. Como el objetivo principal del PCA es la compresión de los datos en variables ortonormales, generalmente no se suelen extraer todas las componentes principales por lo que se concurre en un error.

Supongamos que se extraen m componentes principales, por consiguiente la base de datos original podría recuperarse mediante $\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}$ donde $\mathbf{E} \in \mathcal{M}_{N \times K}(\mathbb{R})$ es la matriz que recoge los errores cometidos al estimar a \mathbf{X} como $\hat{\mathbf{X}}$ mediante m componentes principales.

Realizadas las aclaraciones, cabe manifestar la importancia de la validación del modelo PCA creado antes de presentar sus conclusiones, puesto que podría haberse recogido ruido en el modelo que altere los resultados.

Para ello, se comienza realizando un diagnóstico sobre la bondad de ajuste del modelo; es decir, se evalúa el ajuste del modelo a los datos con m componentes. Dicho diagnóstico se realizará mediante la R^2 , la variabilidad en \mathbf{X} explicada por el modelo de m componentes.

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^N \sum_{j=1}^K e_{ij}^2}{\sum_{i=1}^N \sum_{j=1}^K x_{ij}^2} \quad (3.1)$$

En ciertos puntos de la validación puede resultar de gran interés analizar la variabilidad explicada de cada variable por el modelo de m componentes. Se define para la variable k :

$$R_k^2 = 1 - \frac{SCR_k}{SCT_k} = 1 - \frac{\sum_{i=1}^N e_{ik}^2}{\sum_{i=1}^N x_{ik}^2} \quad (3.2)$$

Por tanto, queda claro que la primera decisión a tomar sobre el modelo debe ser el número de componentes principales a elegir para poder comenzar con la validación. Esta no se trata de una tarea sencilla y diferentes autores han realizado trabajos proponiendo técnicas para realizarla [Camacho and Ferrer, 2014; Josse and Husson, 2012].

La validación cruzada (conocida como *cross-validation* en inglés) es una de las técnicas más extendidas para este cometido. Se trata de extraer cierto grupo de observaciones (en el caso de eliminar una sola, hablaremos de *leave-one-out*), ajustar el modelo a las observaciones restantes y predecir el valor de las observaciones que se dejaron fuera. Haciendo uso de la medida PRESS (3.3), la suma de cuadrados de los errores de predicción, se podrá hacer una idea del error de predicción para el distinto número de componentes principales.

$$PRESS(m) = \sum_{i=1}^N \sum_{j=1}^K (\hat{x}_{ij}^{(m)} - x_{ij})^2 \quad (3.3)$$

Notese que $\hat{x}_{ij}^{(m)}$ hace referencia al valor predicho para el individuo i para la variable j (x_{ij}) mediante m componentes.

Mediante dicha medida, podrá calcularse la Q^2 el cual indicará la bondad de predicción del modelo para las m componentes. Valores cercanos a 1 indican un poder de predicción excelente por el modelo. Es importante analizar la Q^2 puesto que podría dar evidencias de sobreajuste en el modelo.

$$Q^2 = 1 - \frac{PRESS(m)}{\sum_{i=1}^N \sum_{j=1}^K x_{ij}^2} \quad (3.4)$$

No obstante, el objetivo de todos los modelos no suele ser la predicción mediante su uso, por lo que la validación cruzada no siempre debe ser el único criterio a tener en cuenta para la elección de la cantidad de componentes a extraer.

Otras reglas a tener en cuenta suelen ser la selección de las componentes mediante el gráfico del codo, la selección de toda componente con valor propio mayor a 1 o incluso la selección de tantas componentes sean necesarias para la obtención de un porcentaje de varianza explicada determinada.

A pesar de que se hayan presentado diferentes criterios, nunca se debe perder de vista la exploración de las componentes para entender los datos y aunque así no lo indicasen los métodos se deberán extraer las componentes que tengan relevancia para el modelo.

Finalmente, el modelo también deberá validarse para observaciones para así verificar que no haya valores anómalos que modifiquen los resultados. Hablamos de **datos anómalos** (*outliers* en inglés) cuando una observación de la base de datos dista mucho del resto de observaciones del conjunto. Es importante su identificación y en ciertos casos su posterior eliminación, puesto que al no pertenecer a la muestra analizada, podrían introducir ruido en los modelos, dar pie a modelos inestables y como consecuencia a conclusiones erróneas. Para esta tarea, se calculan dos medidas de distancia para cada observación: la suma de cuadrados residual (SCR) y el T^2 -Hotelling.

$$SCR_i = e_i^T e_i = \sum_{j=1}^K e_{ij}^2 \sim g \chi_h^2 \quad \text{y} \quad T_i^2 = \sum_{j=1}^m \frac{t_{ij}^2}{\lambda_j} \sim \frac{(N^2 - 1)m}{N(N-m)} F_{m,N-m} \quad (3.5)$$

donde $g = \frac{s_{SCR}^2}{2SCR}$.

Se consideran con un nivel de confianza del 95 %, observaciones extremas a aquellas que $T_i^2 > \frac{(N^2 - 1)m}{N(N-m)} F_{m,N-m}(0,05)$ y observaciones atípicas a las que $SCR_i > g \chi_h^2(0,05)$.

Deberá tenerse especial cuidado con las observaciones extremas, dado que estas son las que resultan más peligrosas para el modelo al romper con las relaciones encontradas. Las atípicas en cambio, tan solo se distancian del resto de observaciones aunque en ciertos casos deberán eliminarse del modelo de igual manera.

Una vez aclarados todos los aspectos teóricos, se introduce como se realizó la validación del modelo PCA de la imputación elegida.

Elegida la imputación a emplear, se efectuó un análisis de valores anómalos puesto que estos individuos podrían alterar la agrupación de los pacientes. Para alcanzar este propósito, se volvió a analizar el PCA haciendo en este caso especial hincapié en la *SCR* y la T^2 -*Hotelling* (para ello se tuvo que usar las funciones *SCR* y T^2 incluidas una vez más en el Anexo C).

Detectados los pacientes que despuntaban en *SCR*, se analizó su contribución a la *SCR* (mediante la función *Contri* del Anexo C) para detectar la causa por las que resultaron observaciones atípicas. En el caso de los pacientes que despuntaban en T^2 -Hotelling, se analizaron los *scores* para detectar la o las componentes por las que dicho individuo despuntó y finalmente se analizaron los *loadings* de dichas componentes para hallar el motivo de que resultasen observaciones extremas.

Finalmente, se eliminaron de la base de datos aquellos individuos que despuntaban en T^2 -Hotelling, aunque no lo hiciesen necesariamente en *SCR*, por ser los que podrían romper las relaciones entre variables encontradas. Sin embargo, los individuos que despuntaban en *SCR* se mantuvieron en la base de datos, puesto que fuera de ser atípicas, no podrían crear el tipo de problema anteriormente comentado. Esta decisión se tomó teniendo en cuenta que los datos que se estaban tratando eran del ámbito médico, la decisión no hubiera sido la misma de tratarse de una base de ámbito controlado, como los datos de un diseño de experimentos. Además, la decisión de excluir o no pacientes anómalos fue siempre consensuada con el oncólogo.

3.2.3. Clustering

El interés del presente trabajo reside en definir una nueva clasificación para aquellos individuos que no encajan bien en los grupos etiopatogénicos bien conocidos, es decir, los clasificados hasta el momento como *Non-risky* o *No aplicable* y que en adelante llamaremos grupos 'indefinidos'. Para ello, se decidió partir la base de datos en dos. Por un lado se tendría la base con grupos etiopatogénicos conocidos y bien definidos (los que ya han sido descritos en la introducción), por otra parte los pacientes con grupos 'indefinidos'.

Con tal de definir una nueva clasificación de los grupos 'indefinidos', se optó por aplicar técnicas de *clustering* y ver así, si la hubiera, una nueva reagrupación de los datos.

El *clustering* es una técnica de aprendizaje no supervisado utilizada en numerosos campos como el *Machine learning*, el reconocimiento de patrones, el análisis de imágenes o la bioinformática [Madhulatha, 2012; Diday and Simon, 1976; Rokach and Maimon, 2005]. Tiene como objetivo la agrupación de observaciones similares entre ellas y desiguales a las que corresponden a otros *clusters*².

En este trabajo, se aplicó una estrategia de *clustering* poco frecuente, pero que en este caso parecía ser la más adecuada para una nueva reagrupación de los pacientes 'indefinidos'. En efecto, se utilizó una especie de '*clustering* supervisado'.

En primer lugar, y haciendo uso de una ventaja de la que no se suele disponer a la hora de aplicar este tipo de metodología no supervisada, se empleó *clustering* sobre los grupos bien definidos. Se habla de ventaja puesto que de antemano se sabe que dichos grupos etiopatogénicos están bien definidos, por tanto si se encontrase alguna metodología que clasificase a los pacientes en *clusters* similares a los grupos etiopatogénicos bien definidos, podría esperarse que también lo haría adecuadamente con los pacientes en grupos 'indefinidos' puesto que los pacientes provienen de la misma población. Es por ello que hemos llamado a esta estrategia '*clustering* supervisado'.

Para aplicar *clustering* sobre los grupos bien definidos, ante todo fue necesaria la elección de una distancia. En este caso, se probaron la distancia euclídea, la de Manhattan y la de Gower como candidatas para esta técnica.

3.2.3.1. Medidas de distancia

En el *clustering* las medidas de distancia toman una gran importancia puesto que estas serán las que se utilizarán para medir la semejanza entre observaciones. Por consiguiente, será importante entender qué es una distancia.

Distancia: Se define como distancia a toda función, para cualquier conjunto de elementos X ,

$$d : X \times X \longrightarrow \mathbb{R}$$

$$x, y \longrightarrow d(x, y)$$

²*Clusters*: Grupos de observaciones.

que cumpla las propiedades:

- $d(x, y) \geq 0 \forall x, y \in X$
- $d(x, x) = 0 \forall x \in X$
- $d(x, y) = d(y, x) \forall x, y \in X$
- $d(x, z) \leq d(x, y) + d(y, z) \forall x, y, z \in X$

A pesar de que haya numerosas medidas de distancia, en este apartado tan solo se introducirán aquellas que vayan a ser utilizadas en el presente Trabajo de Fin de Máster.

Distancia euclídea Es la distancia más simple de calcular y la más extendida. No obstante, algunos autores cuestionan su uso para realizar *clustering* dado que es una distancia sensible a las unidades de medida de las variables y a los valores anómalos. Cálculo:

$$d_{Euc}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^K (x_j - y_j)^2}, \mathbf{x} = (x_1, x_2, \dots, x_K), \mathbf{y} = (y_1, y_2, \dots, y_K) \in \mathbf{X} \quad (3.6)$$

Distancia de Manhattan También conocida como *city block*, permitió el cálculo de distancias entre dos puntos en situaciones más reales como la ruta más corta en ciudades (los edificios no se pueden traspasar para realizar diagonales). Cálculo:

$$d_{Man}(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^K |x_j - y_j|, \mathbf{x} = (x_1, x_2, \dots, x_K), \mathbf{y} = (y_1, y_2, \dots, y_K) \in \mathbf{X} \quad (3.7)$$

Es una de las distancias a considerar puesto que se suele ver menos afectada por las observaciones anómalas.

Distancia de Gower Hasta el momento, todas las distancias introducidas tan solo admiten como entrada variables numéricas, por tanto si la base de datos en cuestión contuviese variables categóricas deberían incluirse como variables *dummies*³. De la misma manera, existen distancias para el caso en el que todas las variables son categóricas; ahora bien, estas no serán introducidas puesto que categorizar variables numéricas puede llevar a pérdida de información.

³Variable *Dummy*: Sirve para identificar categorías o clases a las que pertenecen las observaciones codificándolas con 0 o 1 (no pertenece/perteneces).

A la hora de utilizar bases de datos con tipos de variables mixtas, la introducida por Gower [Gower, 1971] es la más popular. Cabe destacar que la distancia de Gower, cuenta con la posibilidad de que en la matriz de datos haya valores faltantes; por ello, hace uso del parámetro δ_{xyj} para controlar si la observación $\mathbf{x} = (x_1, x_2, \dots, x_K)$ y la $\mathbf{y} = (y_1, y_2, \dots, y_K)$ son comparables para la variable j , que tomaría valor 1 si lo fueran o 0 en el caso contrario. Cálculo:

$$S_{\mathbf{xy}} = \frac{\sum_{j=1}^k s_{\mathbf{xy}j} \delta_{\mathbf{xy}j}}{\sum_{j=1}^k \delta_{\mathbf{xy}j}} \quad (3.8)$$

donde $S_{\mathbf{xy}}$ denota la similitud entre las observaciones \mathbf{x} y \mathbf{y} , siendo $s_{\mathbf{xy}j}$ calculada según el tipo de variable:

- **Variables numéricas:** $s_{\mathbf{xy}j} = 1 - \frac{|x_j - y_j|}{R_j}$, con R_j el rango de la variable j
- **Variables categóricas:** $s_{\mathbf{xy}j} = \begin{cases} 1 & \text{si } x_j = y_j \\ 0 & \text{sino} \end{cases}$
- **Variables dicotómicas:** $\delta_{\mathbf{xy}j}$ y $s_{\mathbf{xy}j}$ se calculan siguiendo la tabla:

		Valores variable j			
Individuo x	+	+	-	-	
	+	-	+	-	
$s_{\mathbf{xy}j}$	1	0	0	0	
$\delta_{\mathbf{xy}j}$	1	1	1	0	

Determinada la similitud entre ambas observaciones, la distancia de Gower se calcula como:

$$d_{Gower}(\mathbf{x}, \mathbf{y}) = \sqrt{1 - S_{\mathbf{xy}}} \quad (3.9)$$

Cabe aclarar que salvo la distancia de Gower, como ya se ha introducido, el resto de las distancias utilizadas no permiten el uso de datos mixtos y requieren variables numéricas. Por tanto, para incluir las variables categóricas de la base de datos, se crearon tantas *dummies* como categorías tuviesen y estas fueron incluidas en lugar de las variables originales. Además, todas fueron centradas y escaladas.

3.2.3.2. Tendencia de agrupamiento

Aplicadas las distintas medidas de distancia, es importante saber si los datos presentan tendencia de agrupamiento antes de aplicar cualquier técnica de *clustering* sobre ellos. Para comprobarlo sobre los pacientes de grupos bien definidos, se analizaron los *heatmaps* y el estadístico de Hopkins.

Heatmaps Como indica [Weinstein, 2008], los mapas de calor (del inglés *heatmap*), son la representación gráfica más popular para comprimir gran cantidad de información en un espacio pequeño y sacar a relucir patrones existentes en los datos. Suelen ser utilizados en diversos campos como, por ejemplo, en el biomédico.

El uso de los *heatmaps* puede ser de gran ayuda a la hora de visualizar rápidamente si los datos son agrupables o no. Analizar la tendencia natural de agrupación de los datos es de vital importancia puesto que al aplicar *clustering* sobre una matriz de datos, siempre se devolverán *clusters*. En caso de que los datos tuvieran una estructura aleatoria o una distribución uniforme, los *clusters* creados podrían ser engañosos y dar lugar a conclusiones erróneas.

En ciertos casos, en los margenes del *heatmap* suelen añadirse árboles de *clusters*, más conocidos como dendrogramas.

Al crear los *heatmaps*, para confirmar la tendencia de agrupamiento, se buscó que los mapas presentasen una clara diferenciación entre grupos.

Estadístico de Hopkins En la sección anterior, se ha introducido como los *heatmaps* pueden resultar de ayuda a la hora de comprender si los datos que se están manejando son naturalmente agrupables o no. Pese a ser una técnica muy extendida, en ciertas ocasiones no es tan evidente determinar si los datos son agrupables o no tan solo visualizándolo. Es aquí cuando el estadístico de Hopkins toma importancia.

En cierto modo el estadístico de Hopkins mide la probabilidad de que la base de datos esté generada por una distribución uniforme [Han et al., 2012]. Para ello, dada una matriz de datos \mathbf{X} , se quiere determinar cómo de lejos está dicha matriz de estar uniformemente distribuida en el espacio de datos. Para su cálculo se seguirán los siguientes pasos:

1. Se seleccionan aleatoriamente m elementos de la matriz \mathbf{D} , $\mathbf{p}_i \in \mathbf{D}$, $i \in \{1, \dots, m\}$

2. Para cada uno de los \mathbf{p}_i elementos seleccionados, se calcula la distancia del vecino más cercano en \mathbf{D} y se guarda como (x_i):

$$x_i = \min_{\mathbf{o} \in \mathbf{D}} (d(\mathbf{p}_i, \mathbf{o}))$$

3. Se simulan m elementos, \mathbf{q}_i , a partir de una distribución uniforme con el mismo rango que los datos originales.
4. Para cada uno de los \mathbf{q}_i elementos, se calcula la distancia del vecino más cercano en \mathbf{D} y se guarda como (y_i):

$$y_i = \min_{\mathbf{o} \in \mathbf{D}} (d(\mathbf{q}_i, \mathbf{o}))$$

5. Se calcula el estadístico de Hopkins, como:

$$H = \frac{\sum_{i=1}^m y_i}{\sum_{i=1}^m x_i + \sum_{i=1}^m y_i}$$

Si la matriz de datos \mathbf{D} se distribuyese uniformemente, $\sum_{i=1}^m x_i$ y $\sum_{i=1}^m y_i$ serían parecidos por lo que H sería 0.5; sin embargo, cuando existe agregación de los datos por naturaleza, $\sum_{i=1}^m y_i$ debería ser bastante mayor que el de un patrón aleatorio por lo que se esperarían valores de H cercanos a 1.

Para garantizar resultados consistentes, [Banerjee and Dave, 2004] sugiere escoger m de tal manera que $m < 0.1n$. A pesar de ello, para ser más precisos con el cálculo de este estadístico, en el trabajo se probaron distintos m y se fijaron distintas semillas aleatorias.

En caso de que se probase la tendencia natural de agrupamiento de los datos mediante las distancias seleccionadas, se estaría en posición de poder aplicar distintos métodos de *clustering* en busca de aquel que mejores resultados presentase (más adelante se introducirá como se midió que método presentaba los mejores resultados).

3.2.3.3. Métodos de clustering

En la actualidad, existen cientos de métodos para realizar *clustering*; sin embargo, en el presente trabajo se introducirán tan solo algunos de los principales métodos. Los algoritmos que se describen pueden dividirse en cuatro categorías [Milligan and Cooper, 1987]: los métodos jerárquicos, los algoritmos de partición (o también conocidos como no jerárquicos), los métodos de *clusters* difusos y los *clusters* basados en densidad.

Métodos jerárquicos Es uno de los métodos más extendidos. Los algoritmos jerárquicos realizan *clusters* sucesivos usando los previamente establecidos. Pueden ser de dos tipos: aglomerativos o disociativos. Estos últimos no se introducirán puesto que no se utilizaron en el presente trabajo.

Los métodos aglomerativos comienzan considerando a cada observación en un único *cluster* y en cada nivel sucesivo, dos *clusters* se fusionan hasta que se cree un *cluster* en el que todas las observaciones estén incluidas. Una de las características de este método es que una vez que dos elementos pasan a formar parte del mismo *cluster* no vuelven a separarse de ahí en adelante.

Para presentar los resultados, el investigador tiene la opción de usar la jerarquía completa o seleccionar un nivel para la selección de un número concreto de *clusters* de interés. Hay ciertos criterios para determinar que dos *clusters* se fusionarían en cada nivel; sin embargo, solamente se mencionarán: método de la mínima variación de Ward, método de *single linkage* (busca la distancia mínima), método de *complete linkage* (distancia máxima) o el método de la distancia promedio (*average*). Para mayor información véase [gal, 2022].

Los resultados de los métodos jerárquicos podrán graficarse mediante un dendrograma.

Métodos de partición Los métodos de partición, producen una partición de los datos en *clusters* disjuntos de tal forma que los individuos pertenezcan a alguno de los *clusters* posibles. Son también conocidos como métodos no jerárquicos puesto que tan solo producen una partición de los datos. Los algoritmos de partición requieren que previamente el investigador especifique el número de *clusters* que va a utilizar, a posteriori las observaciones se van realojando en los diferentes *clusters* hasta conseguir la convergencia. El algoritmo suele determinar todos los *clusters* a la vez. Los algoritmos más conocidos son: *K-means* y *K-medoids*.

K-means asigna cada observación al *cluster* que tenga el centroide⁴ más próximo. El centroide se calcula como la media de todas las observaciones que pertenecen a dicho *clusters*.

En *K-medoids* cada *cluster* está representado por una observación alojada cerca del centroide del *cluster*, el medoide. En este caso, las observaciones son asignadas a aquellos *clusters* que más cerca tengan los medoides. PAM es uno de los primeros y más conocidos algoritmos de *K-medoids*. Cuando la base de datos contiene observaciones anómalas, este

⁴Centroide: Punto equidistante a las observaciones que pertenecen a un mismo *cluster*.

método suele ser más robusto que *K-means*. Sin embargo, ambos algoritmos inicializan los centroides aleatoriamente.

Como mejora a los algoritmos de *K-means* y *K-medoids* en el caso de datos mixtos⁵ surgió el algoritmo *K-prototypes*.

En *K-prototypes* se definen r individuos ficticios como centroides de los grupos (los prototipos), construidos a partir de la media por grupo para las variables continuas y la moda por grupo para variables categóricas. Por lo demás, el funcionamiento es muy similar al de *K-means*.

Métodos de clusters difusos En los métodos anteriormente presentados, las observaciones tan solo eran asignadas a un solo *cluster*. En los métodos de *clusters* difusos o más conocidos como *Fuzzy clustering*, se permite que las observaciones pertenezcan a múltiples *clusters* indicando el grado de pertenencia que tienen a los diferentes *clusters* [Baadel et al., 2016]. Este método permite también identificar las observaciones anómalas puesto que podrían destacar por tener un grado de pertenencia bajo en todos los *clusters*.

Hoy en día existen varios algoritmos para el *Fuzzy clustering*, algunos basados en grafos⁶. Algunos de los métodos más conocidos son: *Fuzzy K-means*, *Possibilistic Fuzzy C-means*, *Overlapping K-means* o *Overlapping partitioning clusters*. No obstante, en este trabajo tan solo se hizo uso del *Fuzzy K-means*.

El algoritmo *Fuzzy K-means* asigna cada observación a múltiples *clusters* con cierto grado de pertenencia. En el algoritmo, los *clusters* se calculan de la misma manera que en *K-means*, mientras que el grado de pertenencia se calcula como la distancia euclídea entre las observaciones en el espacio original.

Aclarar que para potenciar las posibilidades que ofrece el método *fuzzy* (proporciona el grado de pertenencia a cada *cluster*), se decidió observar el grado de pertenencia a cada *cluster* y se asignó a un grupo 'difuso' a todos aquellos individuos cuyo grado pertenencia a su *cluster* no distase más de un 10 % del grado de pertenencia a otros *clusters* (esto fue llevado a cabo mediante la función *fuzzyalgo* también presente en el Anexo C).

Clusters basados en densidad En este trabajo no fueron usados, pero podría considerarse en el futuro el uso de este algoritmo para este tipo de trabajos.

⁵Datos mixtos: Datos con variables tanto numéricas como categóricas.

⁶Grafo: Representación matemática que permite representar las relaciones de los elementos de un conjunto.

Los métodos testados fueron el jerárquico mediante distintos tipos de criterios de unión de *clusters*, *K-means* y *K-medoids* para las 3 distancias anteriormente introducidas. Cabe aclarar que dada la naturaleza del algoritmo *K-means* no pudo aplicarse mediante la distancia de Gower, pero sí para el resto. Además, se aplicaron *K-prototypes* y *Fuzzy K-means*.

3.2.3.4. Determinación del número óptimo de clusters

En la sección 3.2.3.3, se ha introducido cómo los diferentes métodos, requieren en algún punto del algoritmo que se les indique el número de *clusters* que se desean crear mediante ellos. El objetivo de los métodos es la minimización de la varianza intra-cluster y la maximización de la varianza inter-cluster (observaciones cercanas dentro de los *clusters* pero lejanas a otras pertenecientes a otros *clusters*). La correcta elección del número de *clusters* a crear puede ser una tarea difícil de llevar a cabo, por ello se introducen ciertas opciones que pueden ayudar a determinar el número de *clusters*.

Método del codo. Se grafica la suma de cuadrados total intra-cluster para cada número de *clusters* posible (en un rango aceptable). Mediante este método, se busca aquel número de *clusters* k para el que el incremento de k disminuya notablemente la varianza intra-cluster.

Coeficiente de Silhouette. Se desea maximizar la media de los coeficientes de Silhouette, puesto que este mide cómo de buena es la asignación de una observación a su *cluster*. Para ello, compara la similitud de una observación con el resto de observaciones del mismo *cluster* respecto a las observaciones de otros *clusters*. Para la obtención de los coeficientes de Silhouette (S_i), se comienza calculando a_i , la media de las distancias de la observación i a todos los objetos de su *clusters*. Seguidamente, se calcula la media de las distancias de la observación i a todos los elementos del resto de *clusters*, d_{ij} . Se calcula b_i el mínimo entre todas las distancias calculadas en el paso anterior. Finalmente, se calcula S_i como:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad -1 \leq S_i \leq 1 \quad (3.10)$$

Valores negativos en los coeficientes indicarán asignaciones erróneas de las observaciones a los *clusters*. El promedio de los coeficientes de las observaciones de un *cluster* indica la calidad del *cluster* y el promedio de todas las observaciones la calidad global del *clustering* realizado.

Estadístico Gap. Para poder calcular el estadístico de Gap, primero se calcula la varianza total intra-*clusters*, tal que:

$$W_k = \sum_{r=1}^R \frac{1}{2|C_r|} \sum_{\mathbf{x}, \mathbf{y} \in C_r} d(\mathbf{x}, \mathbf{y}) \quad (3.11)$$

donde C_r denota el r -ésimo *cluster* y d hace referencia a cualquier distancia. A continuación, se generan B bases de datos de referencia a partir de la distribución uniforme en el rango de los valores de las observaciones para cada variable. Se les aplica clustering a cada una de esas B bases de datos, para calcular posteriormente la varianza total intra-*cluster* (W_{kb}^*). Finalmente, se calcula el estadístico de Gap:

$$Gap(k) = \frac{1}{B} \sum_{b=1}^B \log(W_{kb}^*) - \log(W_k) \quad (3.12)$$

Se escoge como número de *clusters* al mínimo, k' , tal que $Gap(k) \geq Gap(k+1) - s_{k+1}$ donde $s_{k+1} = \sqrt{1 + \frac{1}{B}} \sqrt{\frac{1}{B} \sum_{b=1}^B (\log(W_{kb}^*) - \frac{1}{B} \sum_{b=1}^B \log(W_{kb}^*))^2}$ denota la desviación estándar.

Los tres métodos anteriormente introducidos son los más utilizados y fueron los empleados para determinar el número de *clusters* a extraer al usar la distancia euclídea y la de Manhattan. Sin embargo, ni el método del codo ni el estadístico de Gap son métodos aplicables al utilizar bases de datos mixtas. En consecuencia, aunque puedan aplicarse al utilizar la distancia euclídea y de Manhattan, no podrán utilizarse en el caso de la distancia de Gower. Es por ello que, en ese caso, hizo falta definir otro enfoque para la elección del número óptimo de *clusters*. En este caso, a parte de utilizar el coeficiente de Silhouette, se utilizaron ciertas métricas que de normal suelen emplearse en la evaluación de la calidad de los *clusters*. Dichas métricas fueron:

Índice de Dunn. Cuanto mayor sea este índice mayor será la calidad del *clustering*. Para calcularlo, se aplica:

$$D = \min_{i=1, \dots, R} \left(\min_{j=r+1, \dots, R} \left(\frac{\min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})}{\max_{h=1, \dots, R} \max_{\mathbf{x}, \mathbf{y} \in C_h} d(\mathbf{x}, \mathbf{y})} \right) \right) \quad (3.13)$$

TD. Este método se aplicó en el caso de *K-medoids*. En este caso, se busca la minimización de la desviación total (del inglés *Total Deviation*). Se calcula mediante (extraído de [Preud'homme et al., 2021]):

$$TD = \sum_{r=1}^R \sum_{\mathbf{x} \in C_r} d(\mathbf{x}, m_r) \quad (3.14)$$

donde m_r hace referencia al r -ésimo medoide, C_r a su respectivo *cluster* y $d(\mathbf{x}, m_r)$ a la disimilitud entre el sujeto \mathbf{x} y el medoide m_r .

TSD. En el caso de *K-prototypes* se consideró la minimización de la suma total de distancias (del inglés *Total Sum of Distances*) entre los individuos y el prototipo de la clase b_r a la que pertenecen. Calculado por:

$$TSD = \sum_{r=1}^R \sum_{\mathbf{x} \in C_r} d(\mathbf{x}, b_r) \quad (3.15)$$

donde la distancia d se define como: $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p (x_j - y_j)^2 + \gamma \sum_{j=p+1}^K \delta(x_j, y_j)$ con δ como la distancia de *Hamming*.

Diversity. Se consideró como métrica adicional para la elección de número óptimo de *clusters* en *K-prototypes*. Se busca maximizar la métrica calculada por:

$$Diversity = \frac{BSD}{TSumD} \quad (3.16)$$

donde *BSD* hace referencia a la suma de distancias entre *clusters* (del inglés *Between Sum of Distances*) y *TSumD* a la suma total de distancias entre todos los individuos. Calculados por:

$$TSumD = \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}), \quad BSD = \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y}) - \sum_{r=1}^R \sum_{\mathbf{x} \in C_r} d(\mathbf{x}, b_r) \quad (3.17)$$

donde la distancia d hace referencia a la anteriormente expuesta, b_r una vez más al prototipo de la clase r y $i, j \in \{1, \dots, R\}$.

Se resumen por tanto las técnicas empleadas para cada técnica de *clustering*. A la hora de aplicar los métodos jerárquicos con la distancia de Gower, fueron el coeficiente de *Silhouette* y el índice de *Dunn* las empleadas para su evaluación. En la aplicación de *K-medoids* para evaluarlo se tuvo que usar el *TD* (función *TD* incluida en el Anexo C) y se tuvo que programar una función propia (*Gower_pam_silhouette* en el mismo Anexo) para calcular el coeficiente de *Silhouette* puesto que las predefinidas en *R* no permiten su extracción al usar esta distancia. Por último, para evaluar *K-prototypes* también ante la problemática para aplicar las metodologías tradicionales, se hizo uso de *Diversity* (función *diversity* del Anexo C) y del *TSD* (función *TSD* incluida en el mismo Anexo).

3.2.3.5. Medidas de similitud entre clasificaciones

Tras llevar a cabo una metodología rigurosa para la creación de *clusters* sobre los pacientes de grupos etiopatogénicos bien definidos, aprovechando la ventaja de saber a qué grupos pertenecen, era de interés comparar los *clusters* obtenidos con los grupos etiopatogénicos originales. De esta manera, se tendría una idea sobre qué combinaciones de distancias y técnicas de *clustering* presentaban resultados más afines a la realidad. Para ello, se emplearon distintas medidas de similitud como el índice *ARI*, *Jaccard* o la *Kappa* de *Cohen*.

Todas las medidas de similitud aquí presentadas toman valores en el rango [0, 1], donde mayores valores representan mayor similitud.

ARI. Es un índice comúnmente utilizado en la comparación de resultados con un criterio externo. Su cálculo viene dado por:

$$ARI = \frac{\sum_{i=1}^R \sum_{j=1}^P \binom{n_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \quad (3.18)$$

donde $t_1 = \sum_{i=1}^R \binom{n_i}{2}$, $t_2 = \sum_{i=1}^P \binom{n_j}{2}$, $t_3 = \frac{2t_1 t_2}{n(n-1)}$, n_{ij} es el número de observaciones del *cluster* j asignado a i y n_i y n_j son el número de observaciones en los *clusters* i y j respectivamente.

Kappa de Cohen. Es una medida del grado de concordancia entre dos evaluadores. Dicha medida, tiene en cuenta la posibilidad de que la concordancia se produzca por casualidad. Definidos p_0 como la proporción de concordancia observada y p_e como la proporción de concordancia dada al azar, se calcula como:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (3.19)$$

Jaccard index. Compara los miembros de dos conjuntos de datos para ver cuáles son comunes y cuáles distintos. Denotando por C al conjunto del clustering creado y a P a la partición externa, se definen los siguientes valores para su cálculo.

- a: Cantidad de observaciones que pertenecen al mismo *cluster* en los dos conjuntos.
- b: Cantidad de observaciones que pertenecen al mismo *cluster* de C , pero a diferente *cluster* en P .

- c: Cantidad de observaciones que pertenecen al mismo *cluster* de *P*, pero a distinto *cluster* en *C*.

$$Jaccard = \frac{a}{a + b + c} \quad (3.20)$$

Una vez analizados los resultados de cada una de las técnicas y distancias utilizadas, se escogieron los tres métodos que mejores resultados presentaron y posteriormente se aplicaron sobre los pacientes de grupos 'indefinidos'. Para la elección del número óptimo de *clusters* a extraer, se aplicaron las metodologías de la sección 3.2.3.4 una vez más. Se escogió como técnica para la creación de *clusters* de los grupos 'indefinidos' aquella que al proyectarla sobre las componentes de un PCA presentase agrupaciones naturales más distinguidas.

Recapitulando, y para que quede totalmente claro al lector, la estrategia global llevada a cabo en este Trabajo de Fin de Máster fue:

1. División de la base de datos en dos. Por un lado se tenían los grupos etiopatogénicos bien definidos y por otro los grupos 'indefinidos'.
2. Aplicación de los distintos métodos y distancias sobre los grupos etiopatogénicos bien definidos.
3. Selección de los tres métodos que mejores índices de similitud presentaron al comparar los *clusters* creados con los grupos etiopatogénicos bien definidos.
4. Aplicación de los tres métodos seleccionados (con sus correspondientes distancias) sobre los grupos 'indefinidos'.
5. Selección del método que presentase mejores resultados; es decir, que al proyectar los *clusters* obtenidos sobre las distintas componentes de un PCA presentasen agrupaciones naturales más distinguidas.
6. Caracterización de *clusters* obtenidos mediante el método seleccionado.

3.2.4. Caracterización de los clusters obtenidos: PLSDA

Una vez seleccionado el método final de *clustering* para los pacientes de grupos 'indefinidos' y creados los *clusters* para dichos pacientes, se aplicó un modelo PLSDA tomando los *clusters* como variable dependiente y como el conjunto de variables independientes,

X, las variables utilizadas para la creación del *clustering* para evaluar qué variables caracterizaban mejor a los *clusters* creados. Para llevarlo a cabo, en primer lugar se aplicó una validación cruzada de 50 repeticiones de *hold-out* repetido usando el 80% de los datos como entrenamiento y el 20% restante como testeo. En este caso, se tomó la tasa de error de clasificación balanceada (del inglés *Balanced Error Rate*, BER) como medida para la elección del número de componentes PLSDA óptimo a extraer. Los resultados fueron presentados en un *boxplot* y se aplicó un ANOVA para comprobar si había diferencias estadísticamente significativas en la tasa de error de clasificación para los distintos números de componentes extraídas. Si las hubiera, se aplicaría un t-test sobre la tasa de error de clasificación para los dos números de componentes con resultados más afines. Elegido el número de componentes a extraer, finalmente se creó el modelo PLSDA.

El análisis discriminante de mínimos cuadrados parciales (*Partial Least Square Discriminant Analysis*, más conocido como *PLSDA*) es una técnica multivariante supervisada que puede ser usada con propósitos tanto descriptivos como predictivos así como para la selección de variables discriminantes [S. Punla et al., 2022]. El PLSDA, al igual que el PCA, proporciona una estrategia de reducción de la dimensionalidad, por lo que también permite el tratamiento de bases de datos grandes, altamente correlacionadas y en las que se consta de más variables que de individuos.

El objetivo principal de esta técnica es la discriminación entre clases. Para conseguirlo, se busca la proyección de los datos sobre ejes que discriminan entre grupos. Al igual que en el PCA, los ejes que se buscan son ortogonales entre ellos. La principal diferencia con la técnica comentada en la sección 3.2.2 es que en este caso se busca aquel subespacio de proyección que maximice la covarianza entre las variables dependientes y las variables independientes. Por tanto, en este caso se hablará de componentes PLS.

En adelante se explicará el funcionamiento de los modelos PLS puesto que el PLSDA es la versión para variables categóricas del PLS.

En los modelos PLS, de manera similar a los modelos PCA, se contará con una matriz de proyecciones de las observaciones sobre las componentes (los llamados *Scores*), pero en este caso una por cada matriz de datos. Es decir, una para las variables dependientes y otra para las independientes. Retomando la nomenclatura de la sección 3.2.2, $\mathbf{T} \in \mathcal{M}_{N \times K}(\mathbb{R})$ para la matriz de datos $\mathbf{X} \in \mathcal{M}_{N \times K}(\mathbb{R})$ y definimos $\mathbf{U} \in \mathcal{M}_{N \times J}(\mathbb{R})$ para la matriz de las variables dependientes $\mathbf{Y} \in \mathcal{M}_{N \times J}(\mathbb{R})$.

Recordando la ecuación A.1, hacía falta una matriz de *Loadings* para obtener esos *Scores*; sin embargo, en PLS al hablar de las direcciones que maximicen la covarianza no habla-

remos de *Loadings* sino de *Weigthings*. Por tanto, se reescribe la ecuación A.1 como:

$$\mathbf{T} = \mathbf{W}\mathbf{X} \quad (3.21)$$

para la matriz de *weightings* $\mathbf{W} \in \mathcal{M}_{N \times N}(\mathbb{R})$. Y análogamente, $\mathbf{U} = \mathbf{C}\mathbf{Y}$ para $\mathbf{U} \in \mathcal{M}_{N \times N}(\mathbb{R})$.

A la hora de extraer cada componente PLS, se busca la maximización de la covarianza, se podría por tanto reescribir el objetivo como (para la componente i -ésima):

$$\max Cov(t_i, u_i) = \max \sigma_{t_i} \sigma_{u_i} Corr(t_i, u_i) \quad (3.22)$$

De la ecuación (3.22) se concluyen por tanto tres objetivos simultáneos del método PLS:

- Maximizar la variabilidad de la matriz de datos \mathbf{X}
- Maximizar la variabilidad de la matriz respuesta \mathbf{Y}
- Maximizar la relación entre la matriz de datos \mathbf{X} y la matriz respuesta \mathbf{Y}

En conclusión, en comparación con el PCA se sacrificará parte de la explicación de la variabilidad de \mathbf{X} con tal de conseguir la mayor correlación con \mathbf{Y} .

3.2.4.1. Validación del modelo PLSDA

Tras la creación del modelo PLSDA es importante validarla antes de sacar conclusiones a partir de él. Para ello, se utilizaron una vez más la SCR y la T^2 -Hotelling aunque en este caso no se eliminará ninguna observación por mucho que despidan en alguna de ellas.

3.2.4.2. Selección de variables del modelo PLSDA

A continuación, será de gran interés identificar aquellas variables que menos aporten a la creación del modelo para su posterior eliminación del mismo. Para llevar a cabo dicha identificación, se combinó el uso del método VIP (del inglés *Variable Importance in Projection*) y la significación estadística de los coeficientes de regresión. Para ser concretos, en este trabajo se eliminó toda variable que presentase un VIP inferior a 0.8 y/o un p-valor no significativo para los coeficientes de regresión de todas las componentes extraídas en el modelo. Dado que la función de *R* utilizada no proporciona los p-valores de los coeficientes de regresión, se obtuvieron mediante la función programada *p.coef* (véase Anexo C).

El VIP es una medida acumulada que describe la importancia de cada variable de la matriz de datos para explicar \mathbf{Y} . Para una variable k su cálculo viene dado por:

$$VIP_k^2 = n \frac{\sum_{a=1}^m w_{ak}^2 SCEY_a}{SCTY} \quad (3.23)$$

donde $SCEY$ hace referencia a la suma de cuadrados explicada, $SCTY$ a la suma de cuadrados total de \mathbf{Y} , m al número de componentes extraídas por el modelo y w_{ak} al *weighting* de la variable k en la componente a -ésima .

3.2.4.3. Caracterización de los clusters

Obtenido el modelo final (validado y cribado para variables), los resultados fueron presentados mediante la gráfica de los *weightings* sobre las componentes PLS y se extrajeron las conclusiones sobre cada uno de los *clusters* gracias a ella. Para corroborar las conclusiones extraídas, se realizó un análisis univariado para comprobar la dependencia de las variables y los *clusters*, presentando también las tablas de proporciones que muestren las relaciones presentes o las diferencias entre los grupos si se encontrasen. En este análisis se añadió la variable Elastosis que se había mantenido fuera del estudio por ser una reclasificación del CSD, pero era de gran interés para el experto. Dichos análisis univariados fueron llevados a cabo mediante el test de la χ^2 en el caso de variables categóricas y mediante un ANOVA en caso de que las variables numéricas lo permitiesen. Las variables numéricas permitirán la aplicación de ANOVA en caso de que sigan una distribución normal. Para comprobarlo, se utilizó el test de Shapiro-Wilks. En caso de que no se distribuyesen normalmente, se optó por categorizarlas y aplicar un test de independencia. Para ello, se creó una variable categórica con niveles que representaban los distintos rangos de la variable original. Todos los tests realizados fueron posteriormente corregidos por el método de ajuste de Bonferroni [[Weisstein, 2004](#)].

Para finalizar la caracterización de los *clusters*, se realizó otro análisis univariado, mediante el test de *Fisher* para analizar la posible dependencia de las variables externas al *clustering* con los grupos etiopatogénicos bien definidos y los *clusters* creados para los grupos 'indefinidos'. En este caso se usó el test de Fisher en vez de la χ^2 dado que esta última no era aplicable al no haber la cantidad de observaciones mínima por niveles necesaria (al menos 5 observaciones por nivel). Una vez más, se presentaron las tablas de proporciones para poner al descubierto las relaciones presentes.

Cabe mencionar que en caso de que hubiese valores desconocidos o no aplicables no fueron tomados en cuenta a la hora de realizar los análisis univariados anteriormente presentados puesto que, el interés recaía en el comportamiento de las variables y no en las causas de que estas tuvieran valores faltantes o no aplicables.

Dado el carácter protector o agresivo de ciertas mutaciones somáticas frente a la supervivencia del melanoma, también era de interés estudiar la relación entre los grupos etiopatogénicos bien definidos y los *clusters* creados con las diferentes mutaciones somáticas que se disponían. Para ello, se utilizó nuevamente el test de Fisher. Cabe aclarar, que en este caso los v.f. no fueron incluidos, por lo demás el método aplicado fue el mismo que en los estudios univariados explicados anteriormente.

3.2.5. Análisis de supervivencia

Para concluir el estudio, se analizó si la supervivencia era distinta para los grupos etiopatogénicos. Para analizarlo, a petición del experto, se excluyeron los individuos de grupos etiopatogénicos Mucoso y Primario desconocido y aquellos que tuvieran Estadio In situ o A distancia, puesto que su efecto es ampliamente conocido en la literatura y así se facilita la interpretación de los resultados.

El análisis de supervivencia [Kleinbaum et al., 2012] es un conjunto de técnicas estadísticas en las que la variable respuesta mide el tiempo desde el comienzo del seguimiento de un individuo hasta que ocurre un evento; como por ejemplo, la muerte, la recaída o la aparición de cierta enfermedad. Es una técnica muy extendida en el ámbito médico en campos como la investigación sobre el cáncer [Rossi et al., 2012; Xi et al., 2006; Nagore et al., 2005], análisis de problemas hepáticos [El-Serag and Everhart, 2002] o cardiovasculares [Maceira et al., 2008].

Dada la naturaleza del origen de los datos, es común que no se conozca con exactitud el tiempo de seguimiento de los individuos, ya sea porque el estudio finalizase sin que se produjera el evento, porque el individuo fuese extraído del estudio, porque el seguimiento se haya comenzado más tarde que la aparición del evento o a causa de la pérdida de seguimiento. En estos casos, se hablará sobre censura de los individuos, ya sea por la derecha o por la izquierda. Las censuras indican que el tiempo hasta el evento es mayor que el tiempo hasta la censura del individuo, pero que por alguna circunstancia no se posee el tiempo hasta el evento.

Se denota por T a la variable aleatoria que mide el tiempo de supervivencia $T \geq 0$ y por t al tiempo específico.

La censura de los individuos se ve controlada por la variable $\delta = \begin{cases} 1 & \text{si } \text{evento} \\ 0 & \text{si } \text{censura} \end{cases}$

Seguidamente, se introducen dos términos necesarios en cualquier análisis de supervivencia y que describen la distribución de T . Uno es la función de supervivencia, $S(t)$ y el otro la función de riesgo, $h(t)$.

La función de supervivencia, proporciona la probabilidad de que un individuo sobreviva más tiempo que un tiempo específico t ; es decir, $S(t) = P(T > t)$. Teóricamente, la función de supervivencia puede ser graficada como una curva suave, dado que $0 \leq t < \infty$. Sin embargo, en la realidad esta suele verse reflejada por un gráfico escalonado puesto que suelen censurarse individuos y porque los estudios no pueden realizarse en un periodo infinito.

La función de riesgo, proporciona el potencial instantáneo por unidad de tiempo de que se produzca el evento, condicionado a que el individuo haya sobrevivido hasta el momento t (se considera un ratio). Matemáticamente, la función de riesgos, se formula como:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (3.24)$$

Ambas funciones están muy relacionadas puesto que,

$$S(t) = \exp(- \int_0^t h(u) du) \quad \text{y} \quad h(t) = - \frac{dS(t)}{dt} \quad (3.25)$$

En conclusión, se podría decir que el objetivo del análisis de supervivencia radica en la comparación de las funciones de supervivencia y de riesgo para distintos grupos de individuos y la evaluación de posibles factores pronósticos.

Como anteriormente se ha introducido, es común que en estudios del ámbito médico se incluyan datos censurados. Ante esta problemática, surge un estimador no paramétrico de las funciones de supervivencia que tiene en cuenta las censuras, las **curvas de Kaplan Meier**.

Estimador de Kaplan-Meier para la curva de supervivencia:

$$\hat{S}(t_{(f-1)}) = \prod_{i=1}^{f-1} P(T > t_{(i)} | T \geq t_{(i)}) \quad (3.26)$$

donde $t_{(i)}$ denota el i -ésimo tiempo de fallo ordenado. Dicha fórmula tiene en cuenta a los

individuos censurados en el tiempo t , puesto que al censurar a algún individuo se extrae del grupo de individuos en riesgo para el tiempo $t + 1$.

Utilizando dicha fórmula podrían graficarse las funciones de supervivencia según ciertos factores para así analizar su impacto sobre la supervivencia. Si las curvas realizadas no se cruzan, se considera que los distintos grupos de individuos (distinguidos según un factor) tienen diferencias significativas en su tiempo a la supervivencia. Sin embargo, se hará uso del **Log-rank test**, un contraste de hipótesis, para comprobar de manera más contundente si los grupos son estadísticamente equivalentes en cuanto a supervivencia se refiere. Dicho estadístico sigue aproximadamente una distribución χ^2 con $G - 1$ grados de libertad, siendo G el número de grupos del factor analizado, que compara como otros muchos estadísticos lo observado frente a lo predicho en los distintos grupos:

$$\chi^2 \sim \sum_{i=1}^G \frac{(O_i - E_i)^2}{E_i}$$

Hasta ahora, tan solo se han considerado los efectos de las variables explicativas de manera univariada en la supervivencia. A la hora de considerar variables explicativas simultáneamente, el **modelo de regresión de Cox para riesgos proporcionales** es el más extendido, también conocido como Cox PH. Es más popular que la regresión logística puesto que utiliza la información sobre la censura de los eventos al contrario que la regresión logística.

El modelo de Cox PH da una expresión para el riesgo, en el tiempo t para un individuo, dados los valores que este posee en las variables explicativas del modelo (y que denotaremos por $\mathbf{X} = (X_1, X_2, \dots, X_k)$) que se usarán para predecir el riesgo. La función de riesgo se calcula como sigue:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^k \beta_i X_i} \quad (3.27)$$

donde $h_0(t)$ hace alusión a la función de riesgo de referencia y β_i estima una medida del impacto de la variable X_i . Nótese que aunque la función de riesgo de referencia sea una función variable con el tiempo, esta no incluye a las variables del modelo y que el riesgo calculado a partir de las variables, no tiene en cuenta el tiempo. Por tanto, tanto la función de riesgo y el riesgo calculado a partir de las variables son independientes del tiempo y cumplen con la hipótesis de riesgos proporcionales. En ciertos casos, algunas variables no cumplen con la independencia sobre el tiempo, en estos casos no se podrá utilizar un modelo de Cox de PH, pero sí que se podrá hacer uso de los modelos de Cox dependientes del tiempo.

Cabe mencionar que los coeficientes β_i son calculados a partir de la maximización de una función de verosimilitud (no será explicado en el presente trabajo puesto que se realiza de la misma manera que para los coeficientes de la regresión logística). Una vez obtenidos, seremos capaces de calcular los cocientes de riesgos o más conocidos como *Hazard Ratios* que al igual que los *Odds Ratios* de las regresiones logísticas nos darán una medida de cómo incrementa o disminuye el riesgo de un individuo según sus variables explicativas. En este caso también, se hará uso del test de Wald para comprobar si las variables explicativas son estadísticamente significativas.

Validación del modelo de Cox PH Como cualquier modelo de regresión, el modelo de Cox PH también está sujeto a ciertas hipótesis que deberán ser verificadas antes de explotar el modelo. Son tres las hipótesis bajo las cuales está construido el modelo.

1. **Hipótesis de riesgos proporcionales.** La primera a verificar y la más importante, por ello da nombre al modelo. Se requiere que el *Hazard Ratio* sea constante en el tiempo, dado que el efecto de las variables predictoras en el *Hazard* se asume que se mantiene constante en el tiempo. Además, los coeficientes de regresión deben ser constantes en el tiempo. Esta hipótesis puede verificarse utilizando los residuos de *Schoenfeld*, siendo la hipótesis nula que el coeficiente de regresión no es dependiente del tiempo. La hipótesis podría también analizarse de manera gráfica utilizando las curvas log-log (no se utilizará en este trabajo).
2. **Linealidad.** Se debe comprobar que las variables predictoras continuas tengan una relación lineal con los residuos. Esta hipótesis puede ser analizada gracias a los residuos Martingale. Los residuos Martingale tienen media 0, por lo que valores cercanos a 1 indican que el individuo tuvo el evento antes de lo predicho y valores cercanos a -1 denotan eventos más tardíos de lo predicho. La gráfica resultante debería representar una línea recta si la hipótesis se cumpliese.
3. **Residuos.** Gracias a los residuos de *Deviance* o a los *dfbeta* se pueden evidenciar individuos anómalos. El estadístico de *dfbeta* estima cambios en los coeficientes de regresión al eliminar los individuos de uno en uno. Su cálculo viene dado por:

$$DFBETAS_{j,i} = \frac{b_j - b_{j(i)}}{\sqrt{s_{j(i)}^2}} \quad (3.28)$$

donde b_j hace referencia al coeficiente de regresión de la variable j -ésima, $b_{j(i)}$ al coeficiente de regresión de la variable j -ésima si la observación i -ésima fuese eli-

minada y $s_{j(i)}^2$ a la varianza del coeficiente de regresión b_j si la i -ésima observación fuese eliminada. Para ambos residuos se esperaría que por simple azar un (α)% de los individuos superase el límite de $z_{\frac{\alpha}{2}}$ para un nivel de confianza de (1- α)%.

En este trabajo, para realizar el análisis de supervivencia, se comenzó generando curvas de Kaplan-Meier para los grupos etiopatogénicos y para las variables que no se han utilizado para la creación de los *clusters* como el Sexo, el Breslow o la Mitosis. A la hora de analizar variables numéricas, no pueden utilizarse las curvas de Kaplan-Meier por lo que se crearon modelos univariados de regresiones de Cox. En este caso será necesario que se cumpla la linealidad de las variables numéricas. Además, se analizaron los p-valores obtenidos del *log-rank test* y se incluyó en el estudio multivariante toda aquella variable que presentase un p-valor menor de 0.1. Cabe mencionar que el Ganglio centinela y el Total de ganglios positivos fueron excluidos del estudio multivariante puesto que tienen una alta correlación con el Estadio y resultaría peligroso incluirlas juntas. Adicionalmente, se incluyeron la Edad, el Sexo y el Estadio como variables de ajuste para garantizar que la relación del grupo etiopatogénico es puramente debido a su efecto y no a causa de ser, por ejemplo, grupos muy marcados por la edad. Para el análisis de supervivencia multivariante, se utilizó la regresión de Cox.

El modelo inicial, fue cribado hasta dar con aquel que contuviese tan solo variables estadísticamente significativas. Para ello, se empleó el método *backward*. Una vez se tuvo el modelo final, se procedió a su validación. Para ello, se analizó la proporcionalidad de los *hazard* mediante los residuos de *Schoenfeld* y que no hubiera individuos anómalos mediante los residuos *deviance* y *dfbeta*.

En caso de que no se validase el modelo a causa de individuos anómalos, estos deberían ser excluidos y se debería repetir el proceso completo para la creación de un nuevo modelo multivariante para garantizar así, que no se hubiera excluido alguna variable a causa de ellos. En caso de que la proporcionalidad de los *hazard* no se cumpliese, se trataría de variables dependientes del tiempo. En esos casos, se analizaría gráficamente el comportamiento de su *hazard* y se crearía una variable que considere las franjas de tiempo en las que los *hazards* presentasen comportamientos diferentes. Finalmente, se volvería a crear el modelo multivariante, pero en este caso considerando de la interacción con la variable de tiempo creada.

3.3. Software

El presente trabajo de fin de máster ha sido desarrollado en el lenguaje de programación *R* v.4.1.0 mediante el entorno de desarrollo integrado *RStudio*. *R* es un lenguaje y entorno para la computación estadística y la obtención de gráficos que contiene programas informáticos para la manipulación de datos, el cálculo y la visualización de las gráficas.

Todos los resultados que se presentan han sido obtenidos mediante un equipo informático que cuenta con sistema operativo Windows 10, con una RAM de 16GB y procesador de onceava generación Intel ® Core TM i7-1165G7 de 2.80GHz.

3.3.1. Librerías de R utilizadas

Para el desarrollo del trabajo, aparte de las funciones y librerías básicas que *R* ofrece, fue imprescindible la utilización de otras librerías más específicas incluidas en los repositorios *CRAN* y *Bioconductor*.

La imputación de datos faltantes fue llevada a cabo mediante el paquete *mice* v.3.13.0 [van Buuren and Groothuis-Oudshoorn, 2011] (disponible en *CRAN*) que imputa datos multivariados incompletos mediante ecuaciones encadenadas. *mice* permite la selección de predictores, la imputación pasiva, la agrupación automática, el post-procesado de los datos imputados, y dispone de herramientas para la selección de modelos y gráficos de diagnóstico.

Los modelos PCA para el diagnóstico de las imputaciones y los modelos PLSDA para la caracterización de los *clusters* se crearon a partir del paquete *roppls* v.1.24.0 de *Bioconductor* [Thevenot et al., 2015]. *roppls* posibilita la creación de modelos PCA, PLS o PLSDA y ofrece múltiples herramientas para la visualización de resultados y la predicción (en los casos de modelos PLS/PLSDA).

Para la aplicación de *clustering* fue necesario el uso de diferentes librerías todas ellas de *CRAN*. En primer lugar se usó el paquete *cluster* v.2.1.2 [Maechler et al., 2021] para la creación de matrices de distancia, la aplicación del método *K-medoids* y el cálculo del coeficiente de Silhouette. Para la creación de *clusters* jerárquicos, se hizo uso de la librería *factoextra* v.1.0.7 [Kassambara and Mundt, 2020]. En el caso de *Fuzzy clustering*, fue necesario el uso de *fclust* v.2.1.1 [Ferraro et al., 2019] la cuál permitió la aplicación del método *Fuzzy K-means*. En último lugar, se utilizó *clustMixType* v.0.2-15 [Szepannek, 2018] que permite la aplicación del método *K-prototypes*.

Los modelos de supervivencia se construyeron a partir de la librería *survival* v.3.2-11 [Terry M. Therneau and Patricia M. Grambsch, 2000]. *survival* permite la creación de regresiones de Cox, da acceso a las herramientas necesarias para la validación de estos modelos e incluso a funciones para la creación de modelos de supervivencia dependientes del tiempo. Sin embargo, esta librería no permite la visualización gráfica de los resultados. Por ello, para visualizar los modelos univariados, se utilizó la librería *survminer* v.0.4.9 [Kassambara et al., 2021] que permite la creación de gráficas de curvas de *Kaplan Meier*.

3.3.2. Desarrollo de funciones necesarias

En esta sección se incluirá una breve descripción de todas las funciones creadas para el correcto desarrollo del presente Trabajo de Fin de Máster. Para facilidad del lector serán incluidas en orden de utilización para el desarrollo del trabajo. Además, en el Anexo C, se incluirá la programación en *R* de todas ellas.

- **dummytovariable.** Esta breve función, y su variante *dummy2tovariable* para variables de 3 categorías, toman como vectores de entrada por un lado las variables *dummies* correspondientes a las categorías de interés y por otro la variable *dummy* correspondiente al valor no aplicable para reconstruir la variable categórica a la que pertenecían.
- **plotloading.** Permite la visualización de los *loadings*, obtenidos mediante un objeto *opls*, en una escala 1:1. Toma como parámetros de entrada el objeto PCA/ PLS/ PLSDA creado por la librería *roppls* y el número de las componentes que se quieren visualizar.
- **R2varcomp.** Permite la creación de gráficos de R^2 explicada por cada variable en cada componente. Toma como parámetros de entrada un objeto PCA creado con *roppls* y la base de datos utilizada para la creación de dicho objeto.
- **SCR.** Gráfica de la suma de cuadrados residual para los individuos incluidos en el modelo PCA, incluyendo en rojo el límite de control para un nivel de confianza del 95 % y en azul para 99 %. Además, devuelve una lista con los individuos que superaron el límite de confianza del 95 % L , con los que superaron el límite $2L$, con la matriz de residuos E y con la suma de cuadrados residual para cada individuo. Toma como parámetros de entrada un objeto PCA creado por *roppls* y la matriz de datos utilizada.

- **T2.** Gráfica de la T^2 -Hotelling para los individuos incluidos en el modelo PCA, incluyendo en naranja el límite de control para un nivel de confianza del 95 % y en rojo para 99 %. Además, devuelve una lista en la que se guardan los individuos que superaron el límite de confianza del 95 % L y los que superaron el límite $2L$. Toma como parámetro de entrada un objeto PCA creado por *ropls*.
- **Contri.** Gráfica de la contribución a la SCR para un individuo concreto. Toma como parámetros de entrada el número del paciente del que se quiere obtener la gráfica, la matriz de residuos E y la suma de cuadrados residual.
- **hopkins.** Esta función permite el cálculo del estadístico de Hopkins para diferentes distancias, semillas y valores de m (el número de individuos a tomar en cuenta para el cálculo del estadístico). Toma como parámetros de entrada los datos a los que se quiere aplicar *clustering*, m, la semilla y la distancia a utilizar.
- **TD.** Realiza una gráfica para la desviación total obtenida mediante diferente número de *clusters* al aplicar el método *K-medoids*. Toma como parámetros de entrada la base de datos sobre la que aplicar *clustering* y el número máximo de *clusters* que extraer.
- **Gower_pam_silhouette.** Realiza la gráfica para el coeficiente de *Silhouette* para el método *K-medoids* al utilizar la distancia de Gower. Toma como parámetros de entrada, la matriz de distancia de Gower y el número máximo de *clusters* a extraer.
- **diversity.** Realiza la gráfica para la diversidad para el método *K-prototypes*. Toma como parámetros de entrada la base de datos sobre la que aplicar *K-prototypes* y el número máximo de *clusters* a extraer.
- **TSD.** Realiza la gráfica para la suma total de distancias para el método *K-prototypes*. Toma como parámetros de entrada la base de datos sobre la que aplicar *K-prototypes* y el número máximo de *clusters* a extraer.
- **fuzzyalgo.** Asigna a un *cluster* difuso los individuos cuya diferencia de probabilidad de pertenencia entre dos *clusters* sea menor a un 10 %. Toma como parámetro de entrada el grado de pertenencia a los grupos, incluido en el objeto de *FKM*.
- **p.coef.** Permite calcular los p-valores de los regresores del modelo PLS o PLSDA mediante técnicas de permutación. Toma como parámetros de entrada un objeto

PLS o PLSDA generado por *ropls*, el número de permutaciones, la matriz de datos utilizada para crear el modelo y la posición en la que se encuentra la variable respuesta en la matriz de datos.

- ***plotweight***. Permite la visualización de los *weightings*, obtenidos mediante un objeto PLS/ PLSDA creado por la librería *ropls*, en una escala 1:1. Toma como parámetros de entrada el objeto *opls*, el nombre de la variable respuesta y las componentes que se quieren visualizar.

CAPÍTULO 4

Resultados

4.1. Limpieza de datos

Como se ha mencionado en la sección 3.1 el conjunto de datos tratado fue una base de datos del ámbito médico que contiene variables clínico-epidemiológicas, clínico-patológicas, histológicas y de mutaciones somáticas para 2304 pacientes. Por su recopilación en la propia consulta médica, contiene v.f. y posiblemente erratas como consecuencia de introducir los datos manualmente. Fue de gran importancia por tanto una buena limpieza de los datos y supuso casi la mitad del tiempo de realización de este trabajo.

Tabla 4.1: Resumen sobre porcentaje de v.f. para los individuos.

Min	1er Cuartil	Mediana	Media	3er Cuartil	Max
0.000	0.000	0.000	6.473	8.108	72.973

En la Tabla 4.1, se presenta un pequeño resumen sobre los v.f. de la base de datos por individuo. Como se puede observar, la mayor parte de los pacientes no contenían gran cantidad de v.f. (así lo muestra el tercer cuartil); sin embargo, algunos presentaban un porcentaje excesivo de ellos.

Siguiendo la metodología expuesta en la sección 3.2.1.2, fueron 227 los individuos eliminados por contener más del 20% v.f. Además, todas las mutaciones somáticas, *Total de ganglios positivos*, *Queratosis seborreicas*, *Angiomas seniles*, *TIL* y *CSD* tuvieron que ser excluidas de la imputación por contener demasiados valores faltantes (dichas variables contenían desde un 24,84 % a un 66,18 % de v.f.).

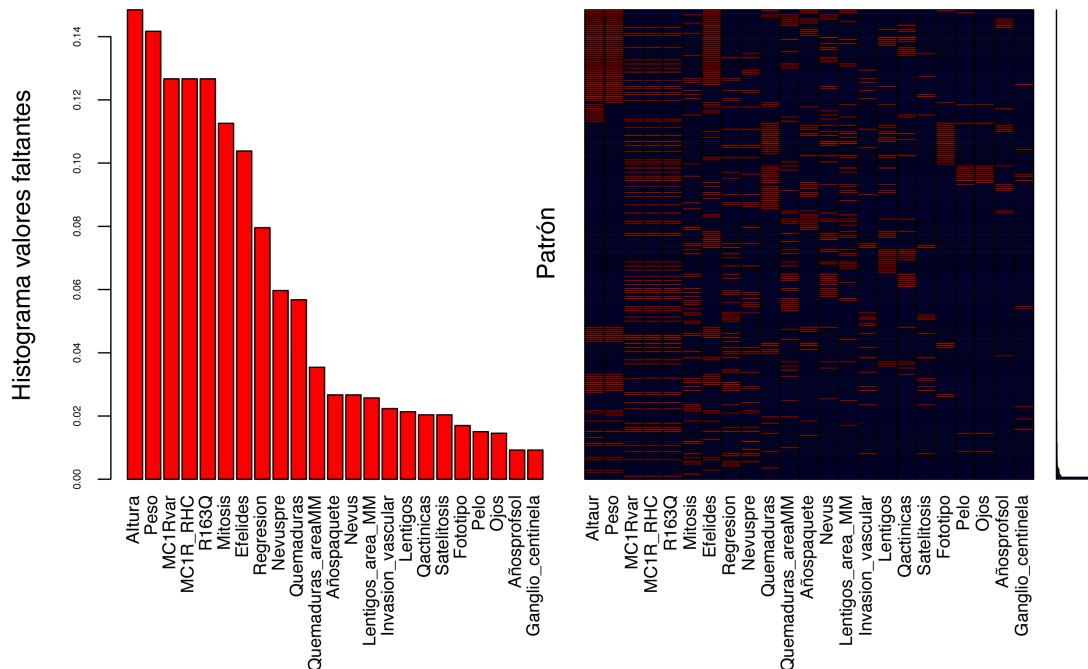


Figura 4.1: Gráfica de co-ocurrencia de v.f.

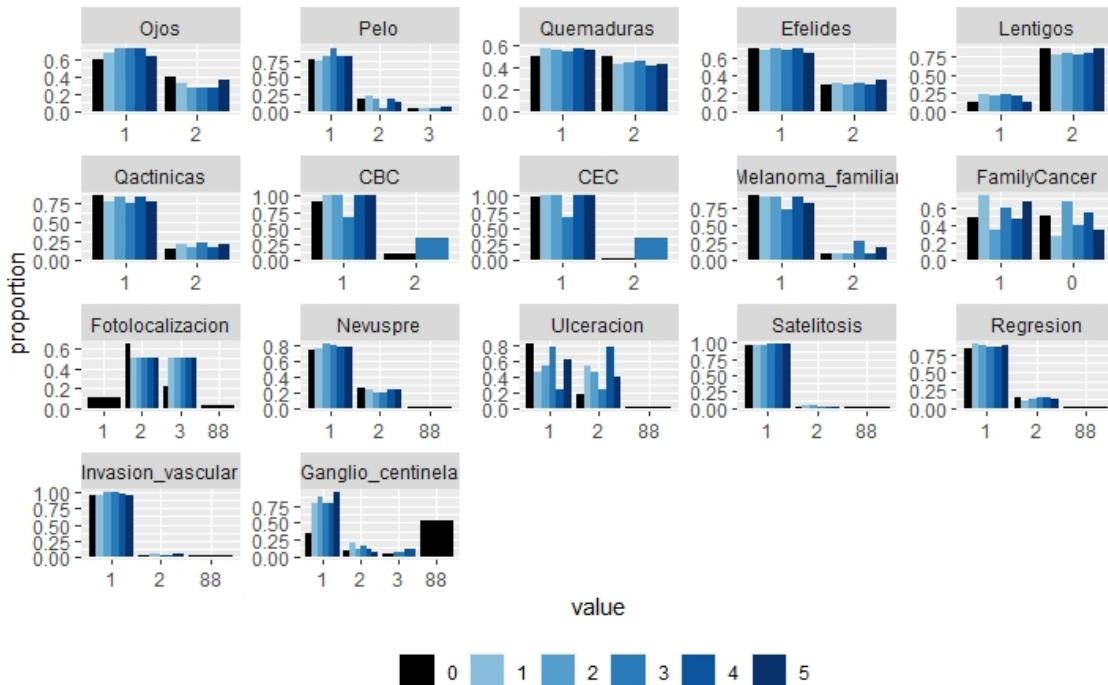
La Figura 4.1 muestra los patrones de aparición de v.f. (se excluyó de la gráfica toda aquella variable que contuviese menos de un 1% de v.f., dado que aunque se encontrase cierto patrón de v.f. al haber tan pocos no se podría tener la certeza de que fuese realmente un patrón o una casualidad y la gráfica se visualiza mejor).

Analizando los patrones de co-ocurrencia encontrados, se pudo observar que aparentemente ciertas variables no presentan v.f. de forma aleatoria, ejemplo de ello son las variables *Altura* y *Peso* o *MC1R*, *RHC* y *R163Q*. El hallazgo no fue inesperado puesto que la relación entre estas variables era conocida de antemano por el experto; no obstante, dichas co-ocurrencias supusieron un reto para la imputación.

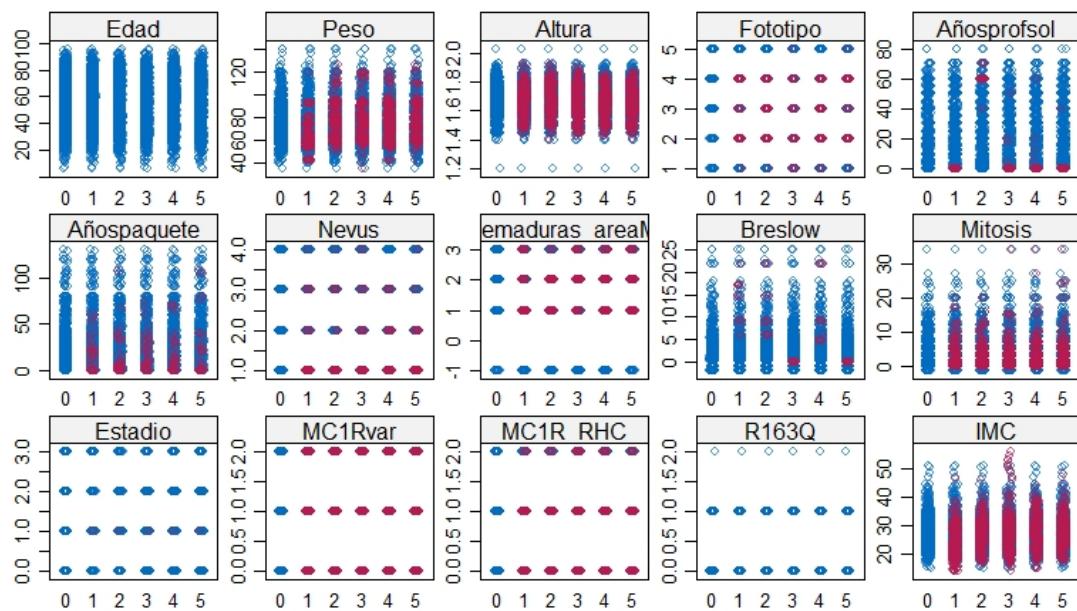
En el caso de las variables *Altura* y *Peso*, no se le dio mayor importancia puesto que el propósito era usar *IMC*. Para imputar esta última, se aplicó la imputación pasiva como se ha indicado en la sección 3.2.1.2 .

El problema fue mayor con las variables *MC1R*, *RHC* y *R163Q*, dado que presentaban un 13% de v.f. y co-ocurrían el 100% de las veces. La imputación de las variables daría lugar a un problema de convergencia, puesto que al imputar una de ellas se utilizaría en la imputación del resto y viceversa. Tras exponer la problemática al experto, se optó por no imputar la variable *R163Q*, la que menos relevancia tiene en este estudio, y se modificaron las matrices de predictores de las otras dos para que no pudiesen usarse como predictoras

entre ellas.

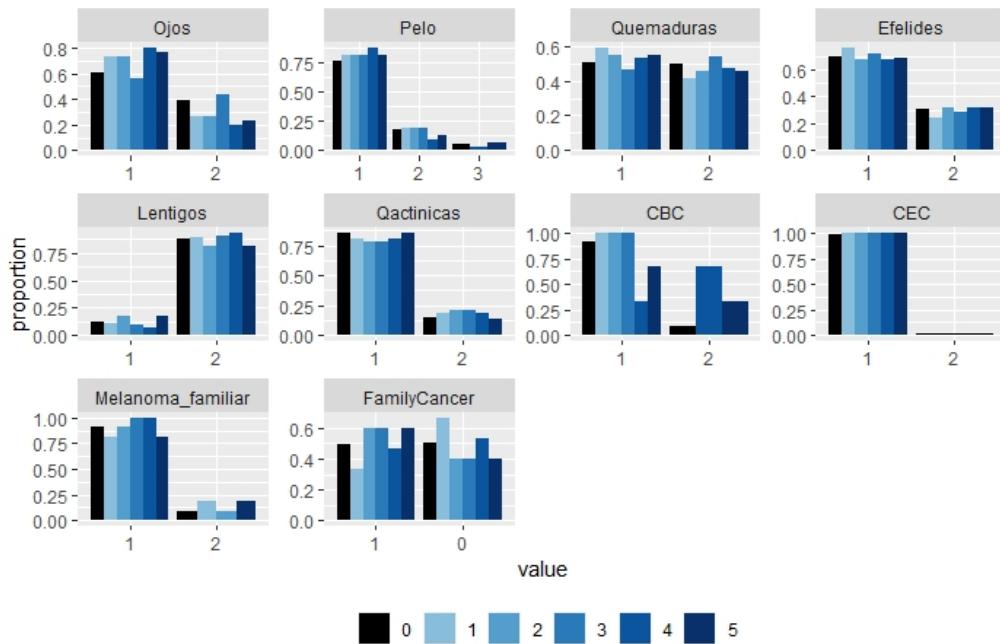


(a) Distribuciones de las variables categóricas imputadas mediante el tipo de imputación 1.

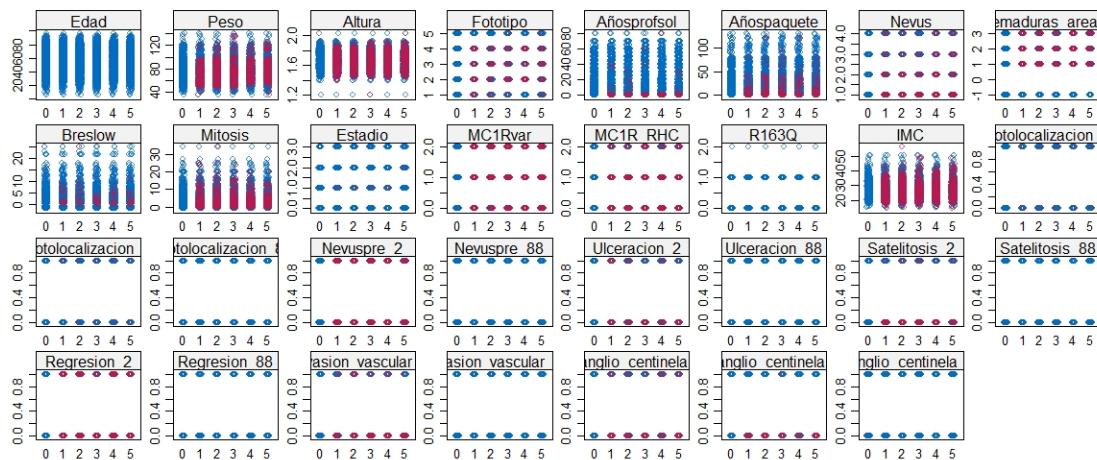


(b) Distribuciones de las variables numéricas imputadas mediante el tipo de imputación 1.

Figura 4.2: Distribuciones para las variables imputadas para el tipo de imputación 1.



(a) Distribuciones de las variables categóricas imputadas mediante el tipo de imputación 2.



(b) Distribuciones de las variables categóricas imputadas mediante el tipo de imputación 2.

Figura 4.3: Distribuciones para las variables imputadas para el tipo de imputación 2.

En las Figuras 4.2 y 4.3, se representan las distribuciones de las variables imputadas atendiendo a las dos opciones de imputación propuestas. Se recuerda que en la primera opción los imputados como 'No Aplicable' (n.a.) se reemplazaron con la clase prevalente y que en la segunda opción las *dummies* correspondientes a n.a. se excluyeron como predictores en el modelo de regresión para evitar su imputación. En este caso, se optó por pedir 5 imputaciones y en las gráficas, se puede observar la distribución que toma cada variable

en cada una de ellas. En ellas también, se pudo observar que la imputación atendió a las restricciones impuestas. En ambos casos, se seleccionó aquella imputación que más se asemejase a la realidad. Escogiendo así la quinta imputación para la primera opción y la segunda imputación para la segunda dado que, en ellas, las variables presentaban las distribuciones más parejas a las de las variables originales sin imputar.

4.2. Validación de la imputaciones mediante PCA

En la Figura 4.4, se presenta un pequeño resumen de los resultados del PCA, para cada uno de los escenarios introducido en la sección 3.2.2, que permitió analizar la calidad de las dos opciones de imputación. Se recuerda brevemente que los escenarios contemplados fueron por un lado la imputación de n.a. mediante la clase prevalente y por otro lado la extracción de las variables *dummies* asociadas a n.a. como predictores del modelo de regresión.

Evaluando la variabilidad explicada por cada modelo (*Explained variance*), la diferencia observada por cada modelo fue ínfima. En el caso del PCA restringido a los datos completos, la variabilidad explicada aumentaba un 0.9 % respecto al resto de modelos; sin embargo, no hay que olvidar que la muestra tuvo que ser restringida a 1190 individuos frente a los 2061 de los que se disponía en los otros casos, por tanto se podía incurrir en pérdida de información.

Examinando el diagnóstico para las observaciones (*Observation diagnostics*), se observaron comportamientos similares en todos los casos. En ellos, se pudo observar que las observaciones que despuntaban en los PCA de los datos originales seguían haciéndolo en los de los datos imputados y que en general no saltaban alarmas por las que pudiésemos pensar que alguno de ellos despuntase a causa de la imputación.

Finalmente, se analizaron tanto los *Scores* como los *Loadings*. En cuanto a los primeros, no se encontraron cambios alarmantes, como mucho un efecto espejo sobre la primera de las componentes. En cuanto al segundo, aunque a simple vista pareciesen bastante parejos, se realizaron los gráficos de los *Loadings* por separado junto con la R^2 explicada por cada variable en cada componente para poder estudiarlos mejor (incluido en el Anexo B.1) y tampoco se encontraron resultados que indicaran que la imputación hubiera modificado la distribución de los datos.

En conclusión, aunque no hubiese evidencias de que una de las opciones de imputación fuese mejor que la otra, se escogió la segunda (la asociada a la creación de *dummies* de

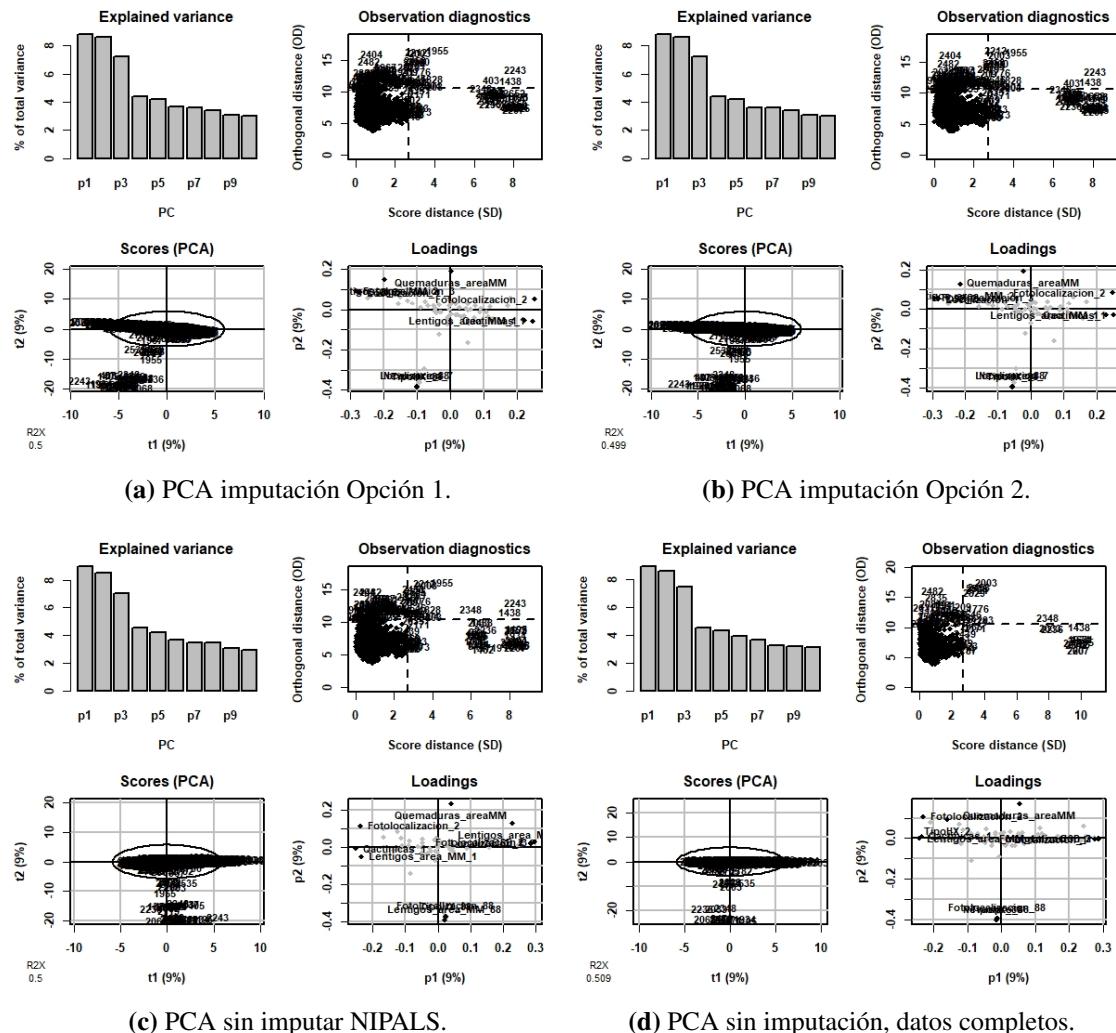


Figura 4.4: Resumen PCA para las bases de datos.

n.a. para su posterior eliminación como regresores en el modelo de imputación) puesto que la primera utilizaba en ciertos casos la imputación de la clase prevalente y esto puede dar lugar a cambios en la distribución de las variables originales.

4.3. Análisis de valores anómalos mediante PCA

Escogida la segunda opción de imputación de valores faltantes como base de datos completa, se procedió al análisis de valores anómalos mediante PCA.

Como se puede constatar en la Figura 4.5, en ambas medidas de distancia, se encontraron observaciones que superaban dos veces los límites de confianza impuestos. En concreto,

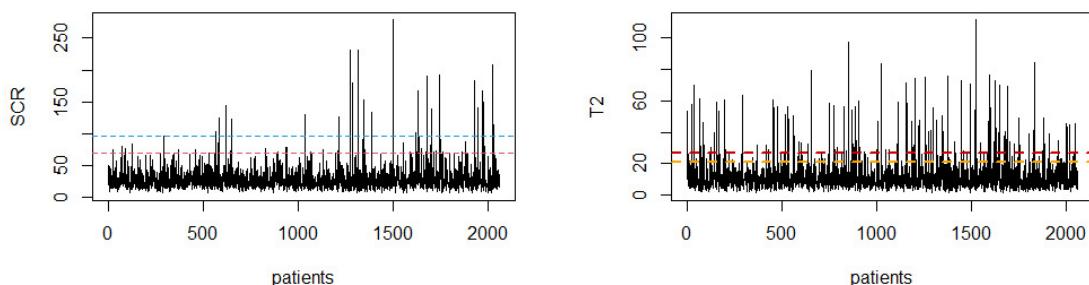


Figura 4.5: Gráficos SCR y T^2 -Hotelling para PCA sobre datos imputados.

se encontraron 15 observaciones atípicas y 55 extremas de este tipo; es decir, observaciones que despuntaban en SCR y T^2 -Hotelling respectivamente. Más allá, el porcentaje de observaciones extremas encontrado (las más peligrosas para el modelo) fue de un 9.9 % que es en cualquier caso superior al 5 % esperado por puro azar; por ello, se tuvieron que analizar las observaciones que despuntaban en esta medida para encontrar la causa de que lo hiciesen.

Como ejemplo del estudio realizado, en la Figura 4.6, se añade el análisis sobre el paciente '1', el primero de dichas características (con el resto se actuó del mismo modo). Estudiando la gráfica derecha de la Figura 4.6, la de los *Scores*, en rojo se pudo apreciar que el individuo '1' se alejaba considerablemente del resto de observaciones en las componentes 1 y 3. Haciendo uso de la gráfica izquierda de la misma figura, se observó que aunque el paciente tuviese lentigos, no tenía lentigos en el área del melanoma yendo así en contra de las relaciones observadas. Además, este paciente contenía quemaduras en el área del melanoma, por lo que se encontró otra ruptura de relación con los lentigos del área del melanoma.

Más allá del paciente '1', se pudo observar un grupo de pacientes que despuntaban por razones similares y que al igual que este paciente pertenecían al grupo de pacientes que podían resultar influyentes para el modelo PCA.

Por otra parte, cabe aclarar que se encontraron ciertos pacientes que siendo de este tipo despuntaron a causa de la localización *Primario desconocido*; sin embargo, no se dispone de muchos individuos de este mismo grupo etiopatogénico y tan solo los hay de esa localización por tanto, se mantuvieron aunque las medidas de distancia recomendadas lo contrario.

Además, dado que se trataron datos del ámbito médico, aunque se encontraron individuos atípicos no fueron eliminados de la base de datos mientras que el experto no lo indicase. A

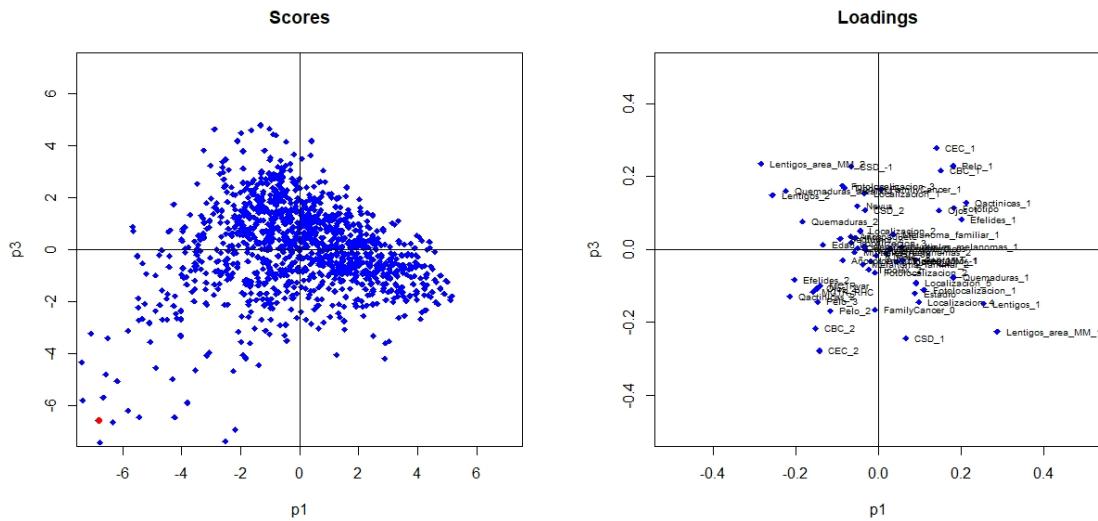


Figura 4.6: Gráficos de Scores y Loadings de las componentes 1 y 3

la hora de analizar los individuos que despuntaron en SCR se hizo analizando las contribuciones del individuo a esa misma medida. Aunque se analizó para todos los individuos atípicos, se incluye tan solo el análisis sobre el paciente '1139' en la Figura 4.7.

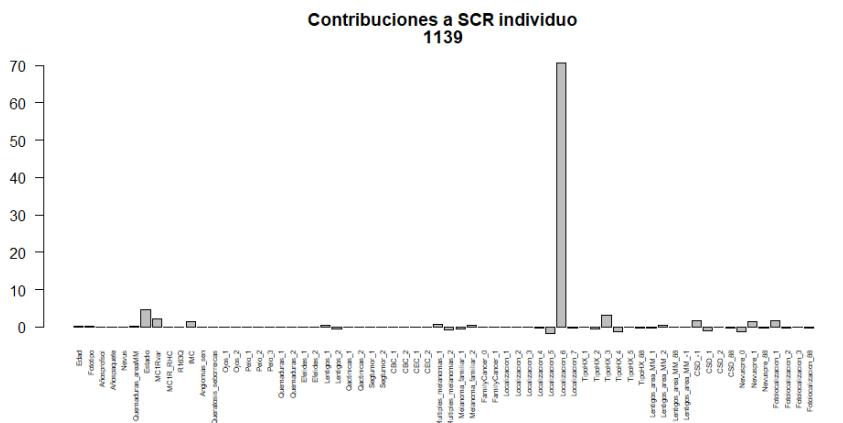


Figura 4.7: Contribuciones a la *SCR* del individuo 1139

Como se puede observar, dicho paciente despuntó por tener el melanoma en una *Localización* poco frecuente: la mucosa. Cabe destacar que todos los pacientes analizados por ser atípicos salvo uno presentaron dicha Localización.

Así pues, fueron 22 los individuos eliminados. Estos pacientes fueron precisamente aquellos que mostraron ser observaciones extremas, pero que no tenían localización *Primario desconocido*. Se tomó la decisión de excluirlos puesto que, aparte de que podrían perjudicar

car los resultados del análisis de *clustering* posterior, no contenían información relevante desde el punto de vista médico. En conclusión, la base de datos final del estudio contenía 2039 individuos para la posterior aplicación del *clustering*.

4.4. Clustering

Una vez depurada la base de datos, se estaba en posición de poder llevar a cabo un análisis *clustering* para la agrupación de los pacientes en grupos etiopatogénicos. En este caso, y como se ha introducido en la sección 3.2.3, se aplicó una especie de '*clustering* supervisado', ya que aunque por definición el *clustering* sea una técnica no supervisada, en este trabajo se pudo aplicar a los pacientes de grupos etiopatogénicos bien definidos con los que se pudo comparar los resultados del *clustering* para identificar los métodos que mejor funcionaban. Se comenzó aplicando una gran cantidad de métodos sobre los pacientes de grupos etiopatogénicos bien definidos y se eligieron las tres metodologías que mejores resultados presentasen basándose en distintas medidas de similitud. Finalmente, se aplicaron esas tres técnicas sobre los pacientes de grupos 'indefinidos' para finalmente escoger aquella que pareciese presentar una mejor agrupación natural y obtener de ella los *clusters* para los grupos 'indefinidos'.

4.4.1. Elección de la distancia y método de clustering

Antes de introducir los resultados, y a modo recordatorio, aclarar que las conclusiones presentes en esta sección corresponden a las obtenidas a partir de la base de datos excluyendo los grupos etiopatogénicos *Non-risky* y *No applicable*; es decir, restringiendo los datos a una muestra de 821 pacientes con grupos etiopatogénicos bien definidos y caracterizados.

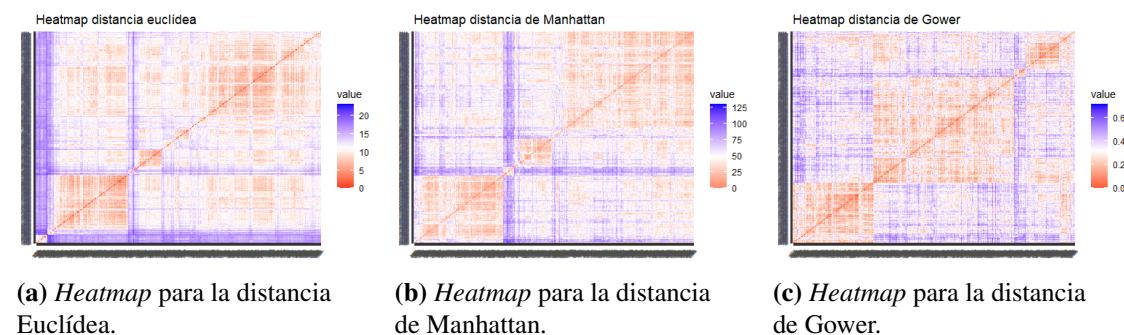


Figura 4.8: *Heatmaps* para los pacientes de grupos etiopatogénicos bien definidos según las distancias.

En la Figura 4.8, se visualizan los *heatmaps* para las tres distancias probadas. Primera-mente, se pudo observar como la distancia euclídea fue la más afectada por las observa-ciones anómalas. Adicionalmente, comentar que los datos sí que daban la impresión de poner al descubierto cierta tendencia a agruparse, pero no todas las distancias parecían presentar el mismo número de clusters. Mientras que las distancias de Manhattan y Go-wer parecían mostrar 4 grupos, en el caso de la distancia euclídea parecían ser 6 grupos, aunque no tan bien definidos.

Tabla 4.2: Resumen estadístico de Hopkins para distintas distancias para grupos etiopatogénicos bien definidos.

Distancia	Min	1er Cuartil	Mediana	Media	3er Cuartil	Max
Euclídea	0.7397	0.7461	0.7476	0.7473	0.7485	0.7507
Manhattan	0.8656	0.8704	0.8713	0.8712	0.8723	0.8745
Gower	0.7569	0.7867	0.8152	0.8134	0.8414	0.8677

Con tal de verificar la tendencia de agrupación visualizada en la Figura 4.8, en la Tabla 4.2 se muestran los resultados obtenidos para el estadístico de Hopkins tras la utilización de diferentes semillas aleatorias y número de m (que tomó los valores: 304, 402, 534, 698). Atendiendo a los resultados aquí expuestos, claramente se pudo concluir que los datos presentaban tendencia de agrupación.

Cabe poner en valor la programación de la función *hopkins*, puesto que obteniendo los mismos resultados mediante ambas, mientras que la función propia era capaz de calcular los resultados entorno a un minuto, la función disponible en *R* para el mismo cálculo tardaba en torno a 15 minutos, y no se podía aplicar a medidas de distancia distintas a la euclídea.

Basándonos en los resultados de la Tabla 4.2 la distancia de Manhattan parecía ser la que mayor tendencia de agrupamiento presentaba y la de resultados más estables; por tanto, una buena distancia para la creación de los *clusters*. Sin embargo, con tal de ser un poco más conservadores, se decidió probar los distintos modelos para las diferentes distancias.

Haciendo uso de la Tabla 4.3, se introduce un pequeño estudio tanto de los métodos de *clustering* como de las medidas de similitud empleados en la aplicación de *clustering* sobre los grupos etiopatogénicos bien definidos.

En la segunda columna de la tabla se añaden los *clusters* extraídos para cada uno de los métodos. En ciertos casos, dicho valor fue extraído automáticamente por las funciones de *R* utilizadas basándose en la optimización de los índices introducidos para la elección del

Tabla 4.3: Resultados de las medidas de similitud para los diferentes métodos de clustering aplicado a individuos con grupos etiopatogénicos bien definidos.

	Método	Número de clusters	ARI	Jaccard Index	ARI 4 grupos	Jaccard Index 4 grupos	Kappa Cohen 4 o 7 grupos
Manhattan	Jerárquico Ward	7	0.5234	0.5009	0.755	0.7809	0.787
	Jerárquico Complete	7	0.4781	0.4819	0.4506	0.8678	0.7655
	Jerárquico Average	4	0.0427	0.189	0.0989	0.3798	0.1397
	Jerárquico Single	4	0.0035	0.1539	0.0064	0.3317	0.0111
	K-means	4	0.5532	0.49	0.827	0.883	0.8805
	K-means	6	0.477	0.47	0.5154	0.357	
	K-medoids	4	0.5147	0.4806	0.7603	0.8408	0.8462
	Jerárquico Ward	4	0.4042	0.4404	0.6667	0.8004	0.838
	Jerárquico Ward	7	0.3996	0.4583	0.6023	0.6338	0.7003
	Jerárquico Complete	4	0.068	0.2029	0.134	0.3998	0.2197
Euclídea	Jerárquico Average	6	0.0457	0.1779	0.103	0.3798	
	Jerárquico Single	3	0.0464	0.2191	0.0969	0.3672	
	K-means	4	0.4867	0.4544	0.696	0.7487	0.7238
	K-medoids	4	0.522	0.4927	0.773	0.8554	0.8671
	Fuzzy K-means	4	0.5047	0.4954	0.765	0.8917	0.9031
	Jerárquico Ward	4	0.5108	0.4766	0.7615	0.8305	0.8454
	Jerárquico Ward	10	0.3374	0.296	0.2969	0.2458	
	Jerárquico Complete	2	0.3036	0.3481	0.5256	0.6355	
	Jerárquico Complete	3	0.3311	0.3857	0.5664	0.691	
	Jerárquico Average	2	0.0451	0.1804	0.095	0.3672	
Gower	Jerárquico Single	2	0.0011	0.1621	0.0026	0.3317	
	K-medoids	3	0.4658	0.4357	0.6881	0.7618	
	K-prototypes	4	0.1801	0.2487	0.3088	0.4229	0.3605

número de *clusters* y en el resto, se seleccionaron siguiendo la metodología presentada en la sección 3.2.3.4 (a modo ilustrativo se incluyen varios ejemplos en el Anexo B.2.1). Como se puede observar en esa misma columna, muchos de los métodos optaron por la extracción de 4 grupos; es más, estudiando los *clusters* extraídos, la gran mayoría de ellos crearon los siguientes grupos: *Nevogénico-Nevogénico débil*, *CSD-Mixto*, *Acral-Mucoso* y *Primario desconocido*. Dicho hallazgo fue comentado con el experto e indicó que concordaba totalmente con lo ya sabido en el ámbito médico, es por eso que en la tabla se añadieron tres columnas adicionales en las que se compararon los resultados del *clustering* con dicha clasificación alternativa basada en los grupos etiopatogénicos bien definidos.

En vista de la Tabla 4.3, los métodos que mejores resultados presentaron fueron el de Ward con la distancia de Manhattan, el de *K-means* para la misma distancia y el de *Fuzzy K-means*. Por ello, fueron estos los métodos que se probarán a posteriori con los pacientes de los grupos 'indefinidos'. En el Anexo B.2.2, puede verse la clasificación que dichos métodos crearon para los pacientes de grupos etiopatogénicos bien definidos. A modo ilustrativo, en el mismo Anexo, se añaden los resultados del método jerárquico mediante

Ward para la distancia de Manhattan, puesto que este fue el que mejor índice de Jaccard presentó y puesto que también era de interés analizar qué clasificación mostraban los pacientes en un *cluster* que había extraído 7 grupos (la cantidad real de *clusters*).

En cuanto a las distancias se refiere, se esperaba que la distancia de Gower fuese la que mejores resultados presentase por tener en cuenta la naturaleza mixta de las variables. Nada más lejos de la realidad. Sin duda alguna esta fue la distancia que peores resultados presentó como bien se puede observar en la Tabla 4.3. Del mismo modo, K-prototypes presentó muy malos resultados aunque este método también tuviera en cuenta la naturaleza mixta de los datos. Siguiendo con el análisis, se pudo observar como la distancia de Manhattan en general presentó mejores resultados que la euclídea en todos los métodos. Para finalizar, se observa que la metodología *Fuzzy K-means* presentó de los mejores resultados de entre todos los métodos.

En cuanto a las medidas de similitud se refiere, cabe comentar que aunque se presenten sus resultados, el índice de Jaccard no dio la impresión de ser la mejor medida de similitud para la toma de decisiones con los datos que se estaban tratando. Esto es así puesto que el índice de Jaccard le da mucha importancia a los casos en los que hay acuerdo entre las dos agrupaciones por tanto, da lugar a índices elevados cuando analizando las agrupaciones creadas se podían observar *clusters* nada claros en los que todos los grupos etiopatogénicos se mezclaban.

4.4.2. Obtención de clusters para pacientes de grupos 'indefinidos'

Entre los métodos comparados en la sección anterior, la única distancia seleccionada fue la de Manhattan, se incluyen por tanto los análisis realizados tan solo con dicha distancia.

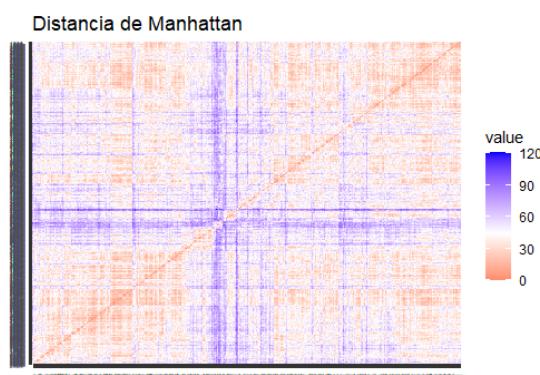


Figura 4.9: Heatmap con la distancia de Manhattan para pacientes de grupos 'indefinidos'.

Analizando la Figura 4.9, se pudo ver como una vez más los datos presentaban cierta tendencia de agrupamiento. Más aún, en este caso parecían crearse 3 *clusters*.

Tabla 4.4: Resumen estadístico de Hopkins para distancia de Manhattan para grupos 'indefinidos'.

Distancia	Min	1er Cuartil	Mediana	Media	3er Cuartil	Max
Manhattan	0.8626	0.8641	0.8648	0.8647	0.8651	0.8668

Una vez más, la Tabla 4.4 fue obtenida tras la aplicación de la función *hopkins* para diferentes semillas aleatorias y número de m (que tomó en este caso los valores: 450, 596, 792, 1035). Observada la Tabla 4.9, se confirmó la tendencia de agrupamiento que podía visualizarse en el *heatmap*.

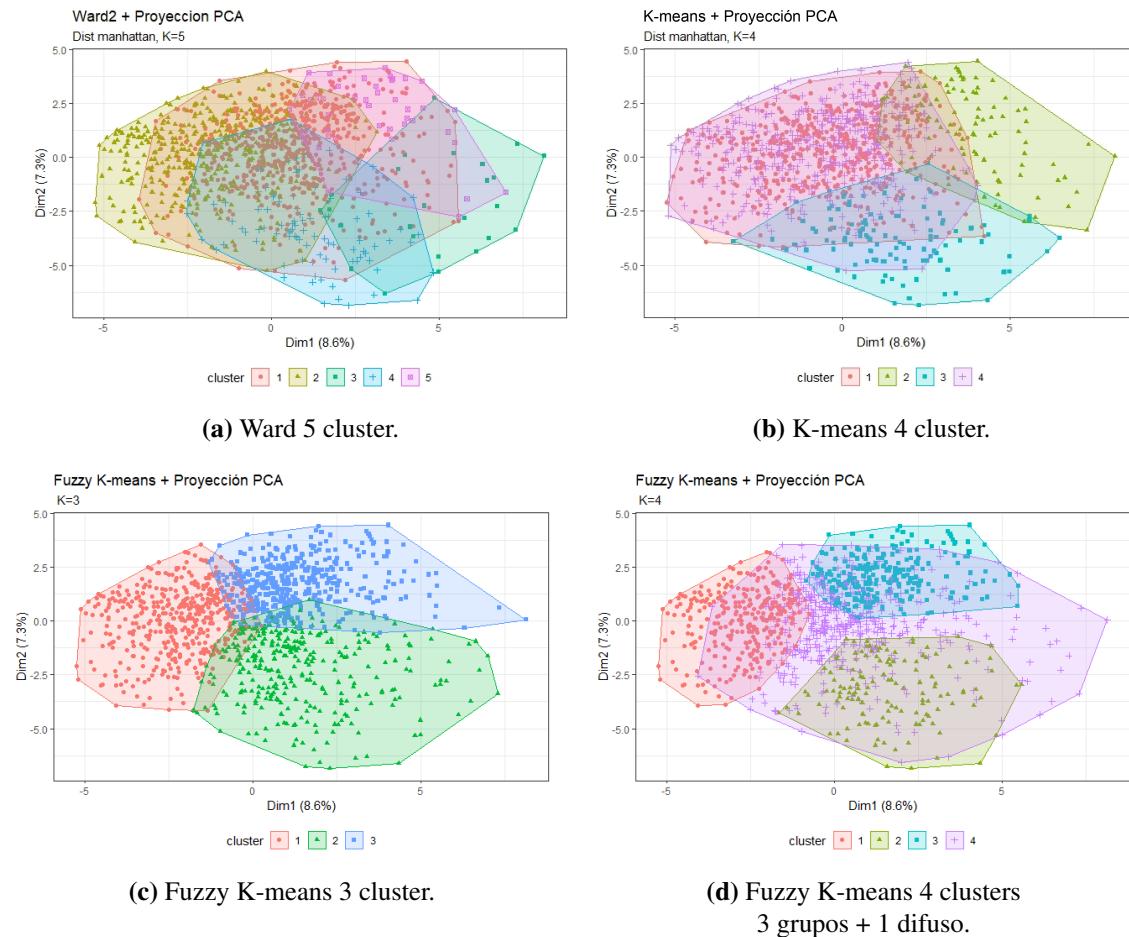


Figura 4.10: Proyecciones de los pacientes de grupos 'indefinidos' sobre las dos primeras componentes principales para métodos escogidos.

En la Figura 4.10, se muestran los *clusters* obtenidos por los métodos de *clustering* prese-

lecccionados en la sección anterior. En el caso de *K-means*, las técnicas de preselección del número de clusters a extraer no dejaron del todo claro cuál debía ser ese valor (incluido en el Anexo B.2.1). La duda surgía entre 4 y 9 *clusters*; sin embargo, como el *heatmap* parecía indicar 3 *clusters*, la extracción de 9 *clusters* no se llevó a cabo por parecer un valor excesivo.

Continuando con el análisis de los *clusters* obtenidos, se observó como los métodos que utilizaron la distancia de Manhattan (el de Ward y el de *K-means*), ya presentaban grupos que se solapaban en las dos primeras componentes principales. En todo caso, el método de *K-means* (Figura 4.10(b)) para 4 *clusters* parecía separar, en cierto modo, los *clusters* 2 y 3 mediante la segunda componente, pero presentaba un solapamiento total entre los *clusters* 1 y 4. Del mismo modo, el método de Ward para 5 *clusters* (Figura 4.10(a)) parecía separar el *cluster* 3 mediante la componente 1, pero no se pudo observar mayor información, ni en las componentes presentadas ni para el resto que se probaron.

En el caso del *Fuzzy K-means* (figuras 4.10(c),(d)), se observaron resultados más prometedores. En la Figura 4.10(c) se incluyen los *clusters* creados obligando a que cada individuo se asigne al grupo de mayor probabilidad de pertenencia; sin embargo, en la Figura 4.10(d), se presentan los *clusters* de los que la pertenencia de grupo está clara más un grupo difuso adicional para aquellos individuos cuyo grado de pertenencia al *cluster* no está clara (siguiendo la metodología que se expuso en la sección 3.2.3.3). Examinando las figuras, en ambas situaciones, se observó como los *clusters* parecían estar mucho más claros que en los casos anteriores en los que todos los grupos se solapaban. Se observó como la componente 1 separaba los *clusters* 2 y 3 del *cluster* 1, siendo la componente 2 la que se encargaba de separar los *clusters* 2 y 3. Por tanto, se podría hablar de *clusters* bien definidos que se separan naturalmente. Anotar que fueron 283 los individuos asignados al grupo difuso de la Figura 4.10(d), grupo que no se diferenció del resto de ningún modo, pero que al crearlo permitió que el resto quedaran mejor definidos.

En cualquier caso, parecía que el método de *Fuzzy K-means* aportaba los mejores resultados y se optó por esta metodología. En cuanto a las dos opciones que este presentaba, por el momento, se mantuvieron las dos y se caracterizaron los tres *clusters* y el grupo difuso obtenidos. En caso de que el grupo difuso no aportase resultados de relevancia se optaría por la opción en la que los pacientes se asignaron obligatoriamente al cluster de mayor probabilidad de pertenencia.

4.4.3. Caracterización de los clusters obtenidos

Con tal de identificar las variables que mejor caracterizaban cada uno de los *clusters* obtenidos, se aplicó un modelo PLSDA sobre cada una de las opciones (los tres *clusters* o los tres *clusters* bien definidos con el grupo difuso) utilizando los *clusters* como variable respuesta.

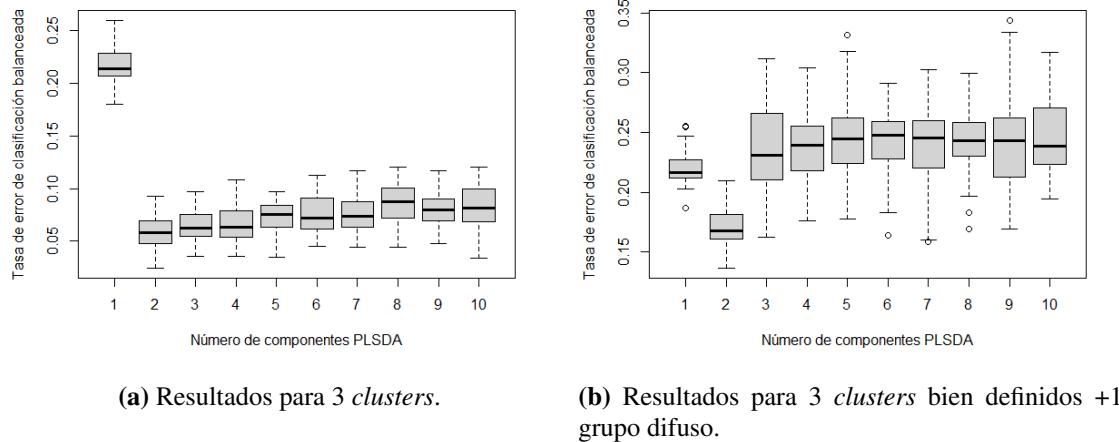


Figura 4.11: Boxplots para la tasa de error de clasificación balanceada para los distintos números de componentes, obtenido por *cross validation*, para las opciones con 3 *clusters* y 3 *clusters* más un grupo difuso.

Analizando la Figura 4.11, se pudo observar claramente como en ambos casos para 2 componentes se hallaba la menor tasa de error de clasificación balanceada en el modelo PLSDA. En ambos escenarios también, el *t-test* confirmó las diferencias estadísticamente significativas entre las tasas de error de 2 y 3 componentes. Por ello, se optó por la extracción de 2 componentes PLS para la identificación de las variables significativas en cada *cluster*. Es de resaltar también las diferencias que se observaron en las tasas de error de clasificación balanceadas entre los dos escenarios. De hecho, como se puede observar en la Figura 4.11, los resultados para 3 *clusters* (Figura 4.11(a)) son en torno a un 10% inferiores que los resultados para 3 *clusters* con el grupo difuso (Figura 4.11(b)).

Ambos modelos fueron cribados utilizando la combinación de VIP mayor a 0.8 y la significancia de los p-valores de los coeficientes de regresión. Tras cribar los modelos, se observó como en el caso de los 3 *clusters* y el grupo difuso ninguna de las variables resultó significativa para este último, como se podía intuir en la Figura 4.10(d). Es por todo ello que se optó por utilizar los 3 *clusters* para la caracterización de los grupos hallados.

Tabla 4.5: Tabla de confusión real frente a predicho del modelo PLSDA creado para la clasificación de los pacientes de grupos 'indefinidos' en los 3 *clusters*.

Predicho	Cluster		
	1	2	3
1	439	21	6
2	1	282	0
3	7	28	434

Analizando la matriz de confusión para el modelo creado (Tabla 4.5), se vio que los *clusters* quedaban casi perfectamente definidos. Lo que corrobora la idoneidad del uso de los 3 *clusters*. Se procedió por tanto a caracterizar los *clusters* por aquellas variables que resultaron significativas y según su posición en los *weightings* de la Figura 4.12.

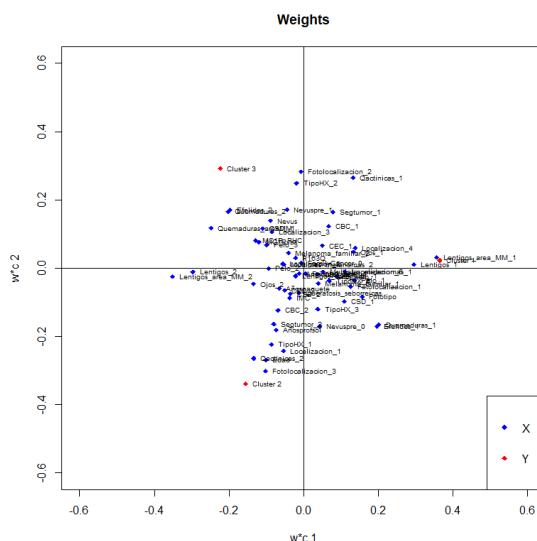


Figura 4.12: *Weightings* de las componentes 1 y 2 para el modelo PLSDA creado.

Cluster 1. Se definiría sobre todo por tratarse de pacientes que no presentan lentigos ni lentigos en el área del melanoma. A su vez, el melanoma se encontró asociado a las localizaciones *Extremidad inferior* y *Acral*. No fue de extrañar por tanto que la fotolocalización *Rara vez* fuese asociada a dicho grupo. Además, se asociaron fototipos altos al grupo, al igual que ojos y pelo oscuros. Tampoco extrañó su asociación con no presentar efélides, ni quemaduras, ni quemaduras en el área del melanoma ni queratosis actínicas debido a su fototipo. Adicionalmente, se trata de un grupo de pacientes que no tiene una edad muy elevada, no ha estado muy expuesto por su profesión a la radiación solar y no suele desarrollar un segundo tumor. Por último, su tipo histológico no suele ser LMM, no suele tener CSD ni MC1R ni RHC.

Cluster 2. Se definiría sobre todo por tratarse de pacientes con edades más avanzadas, fotolocalización *Habitual*, presencia de queratosis actínicas, tipo histológico LMM, no tener tipo histológico MES y presentarse en la localización de *Cabeza/cuello*, pero no en la de *Extremidad* superior ni en la de *Tronco*. Igualmente, suelen ser pacientes que estuvieron años expuestos a la radiación solar por su profesión, pero que no suelen tener historial de quemaduras, que suelen tener pocos nevus, el IMC un poco alto, suelen desarrollar segundos tumores, no tienen restos de nevus en el melanoma y que no suelen tener efélides. De igual manera, se trata de pacientes con ojos más bien claros que en cierto modo tienen lentigos y lentigos en el área del melanoma, además de CSD y CBC.

Cluster 3. Se definiría sobre todo por tratarse de pacientes que suelen tener historial de quemaduras, presentan efélides, tipo histológico MES, fotolocalización *Ocasional* y no presentan queratosis actínicas. A su vez, son pacientes de edades no muy avanzadas que no estuvieron expuestos muchos años a radiación solar por su profesión ni suelen tener fototipos altos, pero que suelen tener recuentos altos de nevus, restos de nevus en el melanoma y presentar quemaduras en el área del melanoma graves. El melanoma suele presentarse en la localización *Tronco*, pero rara vez en la de *Cabeza/cuello*. En cierto modo, este grupo suele presentar MC1R, RHC, asociarse con pelirrojos, tener lentigos y lentigos en el área del melanoma, pero no suele desarrollar un segundo tumor.

En el Anexo [B.3.1](#) se incluyen análisis univariados que apoyaban los resultados aquí presentados.

Expuestos los resultados al experto, indicó que los perfiles de dichos *clusters* eran parecidos a los de grupos etiopatogénicos ya identificados; por tanto, en adelante se hará referencia al **Cluster 1** como *Pseudo Non-risky*, al **Cluster 2** como *Pseudo CSD* y al **Cluster 3** como *Pseudo Nevogénico*.

El hecho de que teniendo perfiles parecidos no estuvieran clasificados desde el principio por los expertos se debía en gran parte porque dichos individuos carecían de ciertas variables que hacen a los expertos tomar la decisión de clasificarlos en unos grupos u en otros.

Como curiosidad, se quiso analizar si en alguno de los *clusters* obtenidos tenía prevalencia alguno de los grupos etiopatogénicos de los que se partía; es decir, los *No clasificable* o los *Non-risky*. Sin embargo, y en contra de lo que se pensaba, en los tres *clusters* se encontró presencia de los dos grupos casi indistintamente (incluido en el Anexo [B.3](#)).

4.5. Relación de los grupos etiopatogénicos con variables de interés

4.5.1. Características histológicas pronósticas

Como se puede observar en la Tabla 4.6, a priori todas las variables presentaron diferencias estadísticamente significativas al analizar su relación con los grupos etiopatogénicos bien definidos y los 3 grupos creados (el *Pseudo Non-risky*, el *Pseudo CSD* y el *Pseudo Nevogénico*). Sin embargo, al aplicar el ajuste de Bonferroni, tanto la *Satelitosis* como el *Total de ganglios positivos* perdieron la significación.

Tabla 4.6: Análisis univariado de las variables externas a la creación de los *clusters* en relación con los grupos etiopatogénicos bien definidos y los grupos creados.

Variables		Tamaño Muestral (n=2039)											
externas	Categorías	Acral	CSD	Mixto	Mucoso	Nevogénico	Nevogénico débil	Primario desconocido	Pseudo Non-risky	Pseudo CSD	Pseudo Nevogénico	p-valor	p ajus
Breslow	[0,1]	0.2247	0.2423	0.2093	0	0.4579	0.4312	0.0294	0.3691	0.2805	0.4966	0.0005	0.005
	(1, 2]	0.191	0.1145	0.2558	0.1579	0.1684	0.1697	0	0.2349	0.1707	0.1959		
	(2,4]	0.191	0.1233	0.186	0	0.1316	0.1468	0	0.1566	0.1402	0.0888		
	[4,∞)	0.191	0.1057	0.093	0.5789	0.0895	0.0963	0.0294	0.1365	0.1982	0.0273		
	In situ (n=387)	0.2023	0.4141	0.2558	0.1053	0.1526	0.156	0	0.1029	0.2104	0.1913		
Mitosis	n. a. (n=35)	0	0	0	0.1579	0	0	0.9412	0	0	0	0.0005	0.005
	0	0.3596	0.5947	0.4651	0.2632	0.4895	0.4037	0.0294	0.3624	0.4024	0.4556		
	[1,5]	0.4607	0.2687	0.4884	0.3684	0.3842	0.4725	0	0.4631	0.4116	0.4806		
	(5, ∞)	0.1798	0.1366	0.0465	0.3684	0.1263	0.1238	0.0294	0.1745	0.186	0.0615		
	n. a. (n=33)	0	0	0	0	0	0	0.9412	0	0	0.0023		
Total de ganglios positivos	0	0.2472	0.1145	0.2326	0.4737	0.2789	0.289	0.1471	0.2774	0.2195	0.1936	0.035	0.375
	1	0.0899	0.0352	0.0233	0.0526	0.0737	0.0872	0.2353	0.1029	0.0854	0.0615		
	[2,3]	0.0785	0.0176	0	0.1053	0.0474	0.0229	0.0588	0.0425	0.0122	0.0137		
	(3, ∞)	0.0225	0.0044	0	0.0526	0.0105	0.0321	0.1176	0.0403	0.0183	0.0091		
	v. p. (n=1302)	0.5618	0.8282	0.7442	0.3158	0.5895	0.5688	0.4412	0.5369	0.6646	0.7221		
TIL	Ausentes	0.5281	0.4361	0.3488	0.6316	0.4737	0.4817	0.0294	0.4318	0.378	0.344	0.005	0.02
	Escasos	0.2584	0.1806	0.4186	0.2105	0.3842	0.4403	0	0.3177	0.3232	0.2141		
	Abundantes	0.0112	0.022	0.0233	0	0.0368	0.0046	0	0.0201	0.0122	0.041		
	n. a. (n=37)	0	0	0	0	0	0	0.9706	0.0022	0.0061	0.0023		
	v. p. (n=517)	0.2022	0.3612	0.2093	0.1579	0.1053	0.0734	0	0.2282	0.2805	0.3986		
Sexo	Hombre	0.4607	0.5418	0.6046	0.2631	0.6	0.5413	0.6176	0.3579	0.6402	0.4624	0.0005	0.005
	Mujer	0.5393	0.4581	0.3953	0.7368	0.4	0.4587	0.3823	0.6421	0.3598	0.5376		
Ulceración	Ausente	0.5843	0.8546	0.814	0.3684	0.8526	0.8394	0.0294	0.7942	0.753	0.9043	0.0005	0.005
	Presente	0.4157	0.1454	0.186	0.6316	0.1474	0.1606	0	0.2058	0.247	0.0957		
	n.a. (n=33)	0	0	0	0	0	0	0.9706	0	0	0		
Satelitosis	Ausente	0.9551	0.9824	1	1	0.9684	0.9725	0	0.962	0.9451	0.9909	0.0165	0.135
	Presente	0.0449	0.0176	0	0	0.0316	0.0275	0.0294	0.038	0.0549	0.0091		
	n.a. (n=33)	0	0	0	0	0	0	0.9706	0	0	0		
Regresión	Ausente	0.9438	0.9427	0.814	0.9474	0.8316	0.8165	0.0588	0.8434	0.8445	0.8041	0.0005	0.005
	Presente	0.0562	0.0573	0.186	0.0526	0.1684	0.1835	0	0.1566	0.1555	0.1936		
	n.a. (n=33)	0	0	0	0	0	0	0.9412	0	0	0.0023		
Invasión vascular	Ausente	0.9663	1	0.9535	0.8947	0.9895	0.9771	0.0294	0.9664	0.9756	0.9886	0.002	0.025
	Presente	0.0337	0	0.0465	0.1053	0.0105	0.0229	0	0.0336	0.0244	0.009		
	n.a. (n=34)	0	0	0	0	0	0	0.9706	0	0	0.0023		
Ganglio centinela	Negativo	0.4382	0.1762	0.3953	0.2105	0.3474	0.3716	0.0294	0.4116	0.3415	0.369	0.0005	0.005
	Postivo	0.1685	0.044	0.0232	0	0.0895	0.1193	0	0.1432	0.1006	0.082		
	No identificado	0.0112	0.0705	0.093	0	0.0316	0.0138	0	0.0291	0.0457	0.0068		
n.a (n=1065)		0.382	0.7093	0.4884	0.7895	0.5316	0.4954	0.9706	0.4161	0.5122	0.5421		

Atendiendo a dicha tabla, se encontraron muchos resultados dignos de mención:

- Breslow: Los grupos que menor Breslow presentaron fueron el *Nevogénico débil* y

los *Pseudo Nevogénico* y *Pseudo Non-risky*. Por el contrario, los *Mucoso* son los que mayor *Breslow* presentaban.

- Mitosis: Los grupos que menor Mitosis mostraron fueron el *CSD* y el *Nevogénico*. En el lado opuesto, una vez más, el *Mucoso* presentó valores más altos que el resto de grupos.
- TIL: En general los grupos presentaron casi a partes iguales *Ausencia* o *Escasos* TIL; sin embargo, en el caso de *Acral*, *Mucoso* y *CSD* la *Ausencia* de TIL se hizo más notoria.
- Sexo: En general se observaron proporcionalmente hombres y mujeres en los diferentes grupos. No obstante, en los grupos *Mixto*, *Primario desconocido* y *Pseudo CSD* la presencia de hombres fue mayor mientras que en los grupos *Mucoso* y *Pseudo Non-risky* las mujeres fueron las prevalentes.
- Ulceración: En general se mostró ausencia de ulceración en los grupos. Pese a ello, en el grupo *Mucoso* la presencia de ulceración fue mayoritaria y en el grupo *Acral* se encontró casi a partes iguales.
- Regresión: En general los grupos mostraron casi por completo ausencia de regresión; sin embargo, en los grupos *Nevogénico débil* y *Pseudo Nevogénico* el porcentaje de ausencia de regresión fue menor que en el resto.
- Invasión vascular: En general los grupos mostraron casi ausencia completa de invasión vascular; no obstante, el grupo *Mucoso* presentó un menor porcentaje de ausencia que el resto.
- Ganglio centinela: Por lo general, los grupos dieron negativo en ganglio centinela; a pesar de ello, los *Acral* y *Pseudo Non-risky* presentaron un mayor porcentaje de positivos en ganglio centinela.

4.5.2. Mutaciones somáticas

En los últimos años múltiples publicaciones han demostrado la relación entre mutaciones somáticas de ciertos genes con la presencia y características del melanoma. En este apartado, se exponen los resultados sobre la asociación de los diferentes grupos etiopatogénicos con la presencia/ausencia de diferentes genes característicos del melanoma.

Tabla 4.7: Análisis univariado de las mutaciones somáticas en relación con los grupos etiopatogénicos bien definidos y los grupos creados.

		Tamaño muestral (n=2039)											
Variables	Categorías	Acral	CSD	Mixto	Mucoso	Nevogénico	Nevogénico débil	Primario desconocido	Pseudo Non-risky	Pseudo CSD	Pseudo Nevogénico	p-valor	p ajust
BRAF	wt	0.8154	0.8	0.8148	0.8	0.5286	0.4803	0.5	0.4886	0.6061	0.5988	0.0005	0.002
	mutado	0.1538	0.1739	0.1852	0.2	0.4714	0.5131	0.5	0.5038	0.3757	0.3827		
	n.a. (n=14)	0.0308	0.0261	0	0	0	0.0066	0	0.0076	0.0182	0.0185		
NRAS	wt	0.8	0.8435	0.8519	0.8667	0.8643	0.9079	0.9091	0.8864	0.8545	0.8312	0.6102	1
	mutado	0.1231	0.1391	0.1481	0.1333	0.1286	0.0789	0.0909	0.0909	0.1212	0.1437		
	n. a. (n=24)	0.0769	0.0174	0	0	0.0071	0.0132	0	0.0227	0.0242	0.025		
KIT	wt	0.7627	0.885	0.85	0.5714	0.982	0.9444	1	0.9005	0.8963	0.9135	0.0005	0.002
	mutado	0.1186	0.069	0.15	0.4286	0	0.0079	0	0.0426	0.0296	0.0096		
	n. a. (n=49)	0.1186	0.046	0	0	0.018	0.0476	0	0.0569	0.0741	0.0769		
TERT	wt	0.9545	0.339	0.5385	0.7778	0.4881	0.5098	0.3333	0.5466	0.4167	0.5217	0.0005	0.002
	mutado	0.0455	0.661	0.4615	0.2222	0.5119	0.4902	0.6667	0.4534	0.5833	0.4783		

En vista de la Tabla 4.7 se pudo observar como todos los genes característicos del melanoma presentaron diferencias estadísticamente significativas en los grupos etiopatogénicos, salvo en el caso del gen NRAS. Se incluyen los resultados observados para el resto de genes:

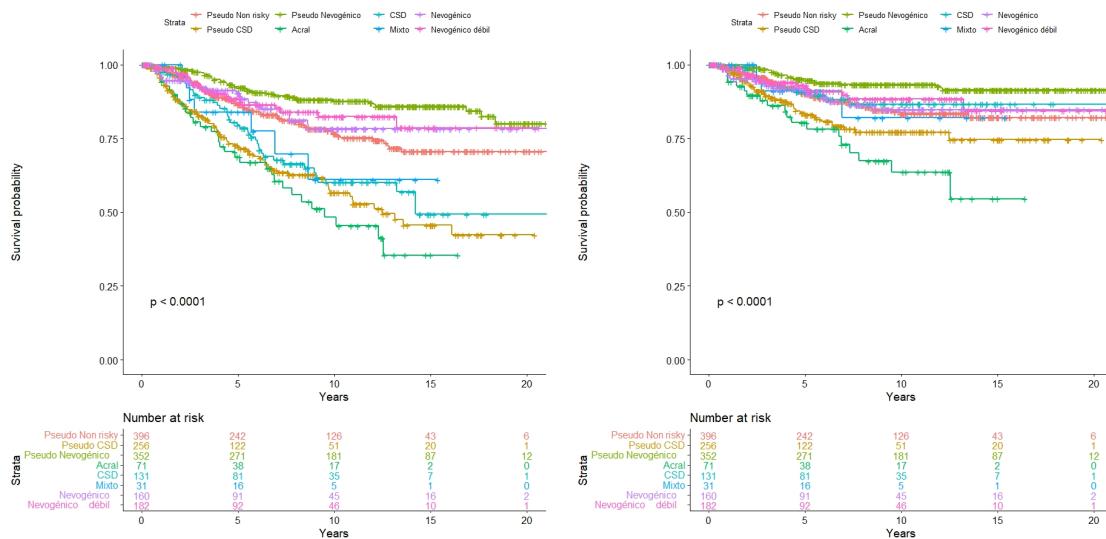
- BRAF: En la mayoría de los grupos etiopatogénicos prima la ausencia de dicho gen; sin embargo, en los grupos etiopatogénicos *Nevogénico*, *Nevogénico débil*, *Primario desconocido* y *Pseudo Non-risky*, la proporción de ausencia y presencia de mutaciones en dicho gen es proporcional. Los casos de *Pseudo CSD* y *Pseudo Nevogénico* presentan un comportamiento un tanto diferente, aunque en ellos prima la ausencia de mutaciones en el gen, se presenta en un porcentaje más elevado de lo común.
- KIT: En el caso de este gen, tan solo se observó un comportamiento diferente en el grupo *Mucoso* y en cierto modo en el grupo de los *Acrales*. Al igual que ocurría en el gen BRAF, prima la ausencia de mutaciones en el gen aunque en el grupo *Mucoso* sea casi proporcional.
- Promotor del TERT: El caso del promotor del TERT es un tanto más complejo. En el caso de los grupos etiopatogénicos *Acral* y *Mucoso*, se da la ausencia de mutaciones en el gen preferentemente. Para los grupos *Mixto*, *Nevogénico*, *Nevogénico débil*, *Pseudo Non-risky*, *Pseudo CSD* y *Pseudo Nevogénico* la ausencia y presencia de mutaciones en el gen es proporcional. Sin embargo, en los grupos *CSD* y *Primario desconocido*, la presencia de mutaciones en el gen es mayoritaria.

4.6. Relación de los clusters con la supervivencia

4.6.1. Curvas de Kaplan-Meier

Antes de mostrar los resultados en relación con la supervivencia, cabe esclarecer que tras eliminar los grupos etiopatogénicos no relevantes (los *Mucosos* y los *Primario desconocidos*), los estadios que no eran de interés (los *In situ* y los *A distancia*) y los individuos para los que no hubo ningún tipo de seguimiento, el tamaño muestral del estudio fue de 1579 pacientes.

Atendiendo a las gráficas de la Figura 4.13, en primer lugar se pudo observar como la supervivencia de los pacientes variaba según si lo que se estudiaba era la supervivencia global o la específica (esclarecer que la supervivencia global hace referencia a muerte por cualquier causa y la específica a muerte específica por melanoma).



(a) Supervivencia global para los grupos etiopatogénicos bien definidos y los grupos encontrados.

(b) Supervivencia específica para los grupos etiopatogénicos bien definidos y los grupos encontrados.

Figura 4.13: Curvas de Kaplan-Meier para el análisis de supervivencia para los grupos etiopatogénicos bien definidos y los grupos encontrados.

En supervivencia global (Figura 4.13(a)), claramente se vislumbró el mal pronóstico de los pacientes de melanoma *Acral* frente al buen comportamiento de los individuos de *Pseudo Nevogénico*. Por otro lado, es de destacar como los grupos que se encontraron en el clustering de la sección 4.4.1 (en este caso: *Acral*, *CSD-Mixto*, *Nevogénico-Nevogénico débil*) presentaron supervivencias parecidas. Es por ello que, y a petición del experto, di-

chos grupos fueron utilizados a la hora de estudiar la supervivencia de manera multivariante.

En supervivencia específica (Figura 4.13(b)), los mejores (los nevogénicos) y peores grupos (los *Acrales* y los *Pseudo CSD*) ante la supervivencia siguieron siendo los mismos; sin embargo, el resto de grupos dejaron de diferenciarse como lo hacían en el caso anterior.

Dicho estudio se reprodujo para todas las variables de interés introducidas en la sección 4.5.1. Como nota aclaratoria añadir que la variable *Mitosis* se incluyó en el estudio de manera categórica desde un principio por tratarse de conteos y que la variable *Breslow* y *Edad* en un principio fueron incluidas como numéricas, pero al no cumplir con el supuesto de linealidad se decidió convertirlas en categóricas siguiendo los cortes propuestos por el experto. En dicho estudio (véase Anexo B.4.1), se concluyó como tanto para la supervivencia global como para la supervivencia específica el sexo femenino, la ausencia de *Satelitosis*, la ausencia de *Ulceración* y la de *Invasión vascular* ofrecían un efecto protector. Por el contrario, se pudo observar como valores elevados en *Breslow*, *Mitosis* o en *Total de ganglios positivos* suponían un mayor riesgo en ambas supervivencias. En el caso de la *Regresión*, tan solo se encontraron diferencias estadísticamente significativas en la supervivencia específica, siendo la presencia de esta la que ofrecía el efecto protector. Finalmente añadir que *TIL* fue la única variable que no presentó diferencias estadísticamente significativas en ninguna de las dos supervivencias.

4.6.2. Modelos de Cox para supervivencia global

Como se incluye en los anexos (véase Anexo B.4.1), en los estudios univariados realizados, *TIL* y *Regresión* (cantidad de células de linfociro infiltrante de tumor y fibrosis en la zona en la que se presentó el melanoma respectivamente) fueron las únicas variables que no presentaron diferencias estadísticamente significativas en la supervivencia global. Sin embargo, y a diferencia de *TIL*, la variable *Regresión* sí que se incluyó en el estudio multivariado por presentar un p-valor poco mayor a 0.05. Además, las variables *Edad* y *Sexo*, se incluyeron estratificadas como variables de ajuste para entender el efecto real del resto de variables incluidas.

En un primer intento, la variable *Estadio* se incluyó como variable de ajuste pero al no cumplir con el supuesto de proporcionalidad de los *hazards* y no ser una variable de especial interés, esta también se incluyó de manera estratificada.

Una vez obtenido el modelo final, se trató de validar el modelo. En la Tabla 4.8, se vio como la *Mitosis* no cumplía con el supuesto de proporcionalidad necesario y por tanto se trataba de una variable dependiente del tiempo. En cuanto a las observaciones anómalas (Figura 4.14), fueron 56 los pacientes que superaron el límite de confianza; sin embargo, se esperaría que 78 pacientes lo superasen por simple azar, por tanto no fue de preocupar.

Tabla 4.8: Residuos Schoenfeld para el modelo multivariante de supervivencia global creado a partir de regresión de Cox.

	p-valor
Etiogrups	0.736
Breslowcut	0.551
Mitosiscut	0.033
Global	0.391

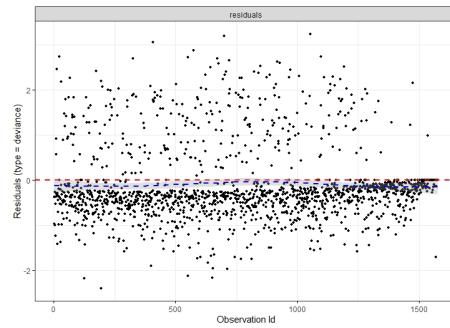


Figura 4.14: Residuos deviance para el modelo multivariante de supervivencia global creado a partir de regresión de Cox.

Con tal de analizar la naturaleza de la *Mitosis* y ver como varían los *hazards* en el tiempo, se estudió la gráfica de la Figura 4.15 donde se pudo observar cierto cambio en el comportamiento de la variable antes y después de 2.8 años de seguimiento. Se utilizó por tanto, dicho punto de corte para la creación de una variable de tiempo que consideraba el tiempo anterior y posterior a 2.8 años que se usaría posteriormente en interacción con la *Mitosis* en un nuevo modelo.

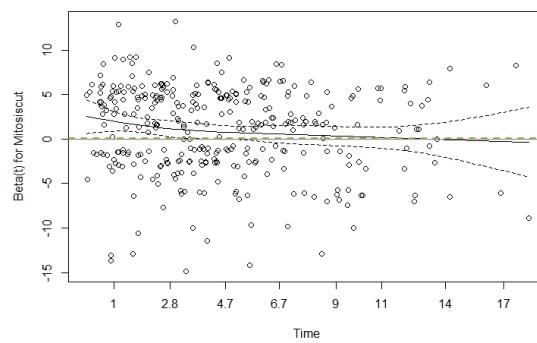


Figura 4.15: Beta dependiente del tiempo para la variable *Mitosis*

En la Tabla 4.9 se incluye el modelo final para la supervivencia global. Dicho modelo, esta vez sí, cumplió con la validación necesaria (Véase Anexo B.4.2).

Tabla 4.9: Modelo de supervivencia global.

Covariata	Categoría	HR	IC 95 % HR	p-valor
Grupo etiopatogénico	Pseudo Nevogénico	Ref	Ref	Ref
	Pseudo CSD	1.697	1.129-2.551	0.0109
	Pseudo Non-risky	1.146	0.77-1.703	0.5022
	Nevogénico-Nevogénico débil	1.207	0.78-1.868	0.398
	Acral	2.193	1.352-3.557	0.0015
	CSD-Mixto	1.379	0.881-2.159	0.1596
Breslow	[0, 1]	Ref	Ref	Ref
	(1, 2]	1.785	1.229-2.59	0.0023
	(2, 4]	2.827	1.926-4.149	$1.11e^{-7}$
	(4, ∞)	3.865	2.581-5.788	$5.31e^{-11}$
Mitosis	0 t.s.<2.8 años	Ref	Ref	Ref
	[1, 5] t.s. <2.8 años	2.473	1.04-5.881	0.0405
	(5, ∞) t.s.<2.8 años	3.97	1.633-9.648	0.00235
	0 t.s. \geq 2.8 años	Ref	Ref	Ref
	[1, 5] t.s. \geq 2.8 años	1.376	0.911-2.081	0.1297
	(5, ∞) t.s. \geq 2.8 años	1.72	1.043-2.837	0.0335

Fueron varios los resultados de interés observados en la Tabla 4.9. Para empezar, se pudo apreciar como tan solo los grupos *Pseudo CSD* y *Acral* presentaron diferencias estadísticamente significativas, en cuanto a la supervivencia global, en comparación con el grupo *Pseudo Nevogénico*. Además, este último como ya se vio en la Figura 4.13 presta un efecto protector, siendo en el caso de los acrales su supervivencia más de dos veces peor. En el caso del *Breslow*, todos los niveles presentaron diferencias estadísticamente significativas en comparación con un valor igual o menor a 1mm de *Breslow*, presentando una supervivencia peor según este incrementa. El caso de la *Mitosis* fue más compleja, en los primeros 2.8 años de seguimiento al tratamiento los dos niveles considerados presentaron diferencias estadísticamente significativas comportándose similar al *Breslow* (a cuanta mayor Mitosis peor supervivencia); sin embargo, pasados los 2.8 años de seguimiento, tan solo tener más de 5 *Mitosis* presentaba diferencias significativas y aunque seguía presentando una supervivencia peor que cuando no se tiene, su efecto era más de 2 veces menor que en el primer tramo de seguimiento.

Ante los resultados observados en la Tabla 4.9 y la Figura 4.13, resultó de interés volver a repetir el estudio de supervivencia pero en este caso juntando los grupos *Pseudo Nevogénico* con los *Nevogénico-Nevogénico débil* y los *Pseudo CSD* con los *CSD-Mixto* puesto que a parte de presentar un perfil parecido (por ello su nombre), presentaban una

supervivencia parecida.

Una vez más se siguió el proceso completo para llegar hasta un modelo estadísticamente significativo (Tabla 4.10) y que cumpliese con los supuestos necesarios para su validación (Véase Anexo B.4.2).

Tabla 4.10: Modelo de supervivencia global juntando los grupos de perfiles etiopatogénicos parecidos.

Covariata	Categoría	HR	IC 95 %	HR	p-valor
Grupo etiopatogénico	(Pseudo) Nevogénico -Nevogénico débil	Ref	Ref	Ref	
	Pseudo Non-risky	1.044	0.752-1.449	0.7957	
	Acral	2.006	1.304-3.085	0.0015	
	(Pseudo) CSD-Mixto	1.429	0.101-1.961	0.0273	
Breslow	[0, 1]	Ref	Ref	Ref	
	(1, 2]	1.772	1.221-2.572	0.0026	
	(2, 4]	2.816	1.922-4.124	1.06e ⁻⁷	
	(4, ∞)	3.878	2.595-5.796	3.83e ⁻¹¹	
Mitosis	0 t.s.<2.8 años	Ref	Ref	Ref	
	[1, 5] t.s. <2.8 años	2.486	1.046-5.905	0.0391	
	(5, ∞) t.s.<2.8 años	4.006	1.651-9.724	0.00216	
	0 t.s. \geq 2.8 años	Ref	Ref	Ref	
	[1, 5] t.s. \geq 2.8 años	1.395	0.925-2.105	0.1122	
	(5, ∞) t.s. \geq 2.8 años	1.72	1.044-2.834	0.033	

Comparando los resultados de la Tabla 4.10 con los de la Tabla 4.9, se puede observar que al juntar todos los pseudo perfiles con los reales, las diferencias de los nevogénicos con el resto de grupos etiopatogénicos se hizo más notoria; no obstante, no se encontraron diferencias estadísticamente significativas con el grupo *Pseudo Non-risky*. No se encontraron mayores diferencias que mínimas variaciones en los *hazard ratios* en comparación con el modelo anteriormente presentado.

4.6.3. Modelos de Cox para supervivencia específica

A la hora de crear el modelo multivariante para la supervivencia específica, una vez más, se comenzó analizando las variables de interés introducidas en la sección 4.5.1 de manera univariada. Como se incluye en los anexos (véase Anexo B.4.1), en el caso de la supervivencia específica la única variable que no resultó estadísticamente significativa de manera univariada fue *TIL*. En este caso, el procedimiento para la creación del modelo

multivariante fue el mismo que el seguido para la creación del modelo multivariante para la supervivencia global de la sección anterior.

Una vez más, se hizo el intento de incluir el *Estadio* como variable de ajuste, pero al no cumplir de nuevo con la proporcionalidad de los *hazards* tuvo que ser incluida de manera estratificada.

El modelo final se incluye en la Tabla 4.11, el cuál fue validado cumpliendo con los supuestos de proporcionalidad para los *hazards* y validando para observaciones anómalas (Véase Anexo B.4.2).

Tabla 4.11: Modelo de supervivencia específica.

Covariata	Categoría	HR	IC 95 % HR	p-valor
Breslow	[0, 1]	Ref	Ref	Ref
	(1, 2]	3.897	2.098-7.238	$1.66e^{-5}$
	(2, 4]	7.192	3.848-13.44	$6.27e^{-10}$
	(4, ∞)	12.593	6.687-23.717	$4.39e^{-15}$
Ulceración	No	Ref	Ref	Ref
	Sí	1.435	1.035-1.989	0.0305

Para finalizar este estudio es de interés analizar los resultados observados en la Tabla 4.11. Para empezar, se pudo apreciar como las variables *Mitosis* y *Grupo etiopatogénico* dejaron de ser significativas. En el caso del *Breslow*, se observó bastante similitud con la supervivencia global en cuanto a comportamiento se refiere; sin embargo, en este caso sus efectos eran mucho mayores (en el caso de tener más de 4mm de *Breslow* su efecto negativo en la supervivencia específica se vio incrementado casi 4 veces). Finalmente, la *Ulceración* resultó estadísticamente significativa presentando una peor supervivencia aquellos individuos que si que presentaban esta variable.

Como se hizo en la sección anterior, también se creó un modelo con los perfiles de grupos etiopatogénicos parecidos juntados. Sin embargo, en este caso no se añadirá el modelo final encontrado puesto que resultó ser el mismo de la Tabla 4.11.

CAPÍTULO 5

Conclusiones

No todos los pacientes de una misma enfermedad se comportan frente a esta de la misma manera ni la desarrollan del mismo modo, es por ello que cobra gran importancia la creación de grupos etiopatogénicos que agrupen a pacientes con características clínicas, epidemiológicas y genéticas similares. Sin embargo, en ciertos casos no está clara la clasificación de los pacientes en dichos grupos etiopatogénicos y es de gran interés la creación de nuevos grupos que los caractericen mejor. Es en este hecho en el que recae el interés del presente Trabajo de Fin de Máster que abordará este problema a partir de una base de datos de unos 2300 pacientes de melanoma del IVO.

A continuación se resumen los principales hallazgos y conclusiones del trabajo realizado:

1. Se ha conseguido crear grupos de pacientes diferenciados según su etiopatogenia. Para su obtención, la metodología llevada a cabo y sus conclusiones fueron:

- Se comenzó imputando los datos faltantes con tal de perder la mínima cantidad de información posible. Para ello, antes de imputar los datos faltantes mediante técnicas basadas en modelos predictivos, se realizó un previo preprocesamiento y limpieza de los datos. A continuación, se analizó y trató la existencia de valores anómalos mediante modelos PCA. Mediante dicha metodología se pudo concluir que se había realizado una minuciosa imputación de la base de datos dado que no presentó diferencias notorias con respecto la naturaleza de la base de datos original y que se obtuvo una base de datos del todo depurada que no produjo problemas en los posteriores pasos del análisis realizado.

- En disposición de una base de datos limpia y completa, se comenzó con la aplicación de técnicas de *clustering* para la búsqueda de nuevos grupos etiopatogénicos para esos pacientes de etiopatogenia 'indefinida'. Con tal de identificar la técnica de *clustering* que pudiese generar los *clusters* para esos pacientes de manera adecuada, primero se aplicaron las diferentes técnicas de *clustering* sobre los pacientes con grupos etiopatogénicos bien definidos y se eligieron las tres técnicas que presentaron los índices de similitud más elevados entre los *clusters* creados y los grupos etiopatogénicos a los que pertenecían. En contra de lo esperado, los métodos de *clustering* específicos para bases de datos mixtas (la naturaleza de los datos de este trabajo), presentaron los peores resultados de entre todos los examinados. El método *K-means* mediante la distancia de Manhattan, el *Fuzzy K-means* y el jerárquico de Ward mediante la distancia de Manhattan presentaron los mejores resultados del estudio. Dichas técnicas se aplicaron posteriormente a los grupos etiopatogénicos 'indefinidos'. Finalmente fue el método *Fuzzy K-means* el utilizado en la generación de los clusters de los pacientes con etiopatogenia difusa por presentar los grupos más diferenciados en las proyecciones sobre las componentes del PCA y ser los que mayor sentido tenían desde el punto de vista médico.

2. Se ha conseguido caracterizar los nuevos grupos etiopatogénicos .

- Elegida la técnica de la cual se trajeron los *clusters* para esos pacientes de etiopatogenia difusa, se caracterizaron haciendo uso de modelos PLSDA tomando como variable dependiente el *cluster* a los que habían sido asignados. Tras la identificación de las variables que los caracterizaban mediante el modelo PLSDA, se ha encontrado como ciertos pacientes presentaban un perfil muy parecido al de algunos grupos etiopatogénicos ya definidos, por ello se les identificó como *Pseudo CSD*, *Pseudo Nevogénico* y *Pseudo Non-risky*. La causa de que no estuvieran ya identificados podía deberse a la falta de valores en las variables decisivas, por lo que el análisis aquí realizado apoya una vez más la calidad de las imputaciones llevadas a cabo y brinda a los expertos una herramienta para clasificar a los pacientes en este tipo de casos. Además, mediante tests univariados se ha confirmado la relación de los grupos y las variables que los caracterizan, así como su relación con variables no utilizadas para la creación de los *clusters*, como lo fueron las variables de características histológicas pronósticas y las de mutaciones somáticas.

- Finalmente, se realizó un análisis de supervivencia con tal de analizar cómo se comportaban los diferentes grupos etiopatogénicos. De este modo se ha encontrado, como ya era conocido por los expertos, el efecto negativo en la supervivencia de las variables *Breslow* y *Mitosis*, siendo el de esta última dependiente del tiempo en el caso de la supervivencia global. Además, también en la supervivencia global, se ha encontrado como el grupo etiopatogénico también toma un papel importante, siendo el grupo *Pseudo Nevogénico* de los que mejor supervivencia presentaba. En el caso de la supervivencia específica, no volvió a encontrarse la importancia de los grupos etiopatogénicos, pero sí la del *Breslow*.

Cabe resaltar el trabajo realizado una vez más dado que, en el presente Trabajo de Fin de Máster, se ha conseguido dar una herramienta alternativa a los médicos para la clasificación de aquellos pacientes con etiopatogenia que hasta el momento se consideraba 'indefinida'. Gracias al desarrollo de este trabajo y la creación de los nuevos grupos etiopatogénicos para los pacientes en esa situación, se ha conseguido dotar a los médicos de una nueva herramienta que podrá clasificar a los pacientes y por tanto permitirles el tratamiento de los pacientes de manera más adecuada según su etiopatogenia.

Anexos

ANEXO A

Material y Métodos

A.1. Descripción de la base de datos

En esta sección se incluye un detallado resumen de las variables utilizadas en el modelo indicando su significado, su nomenclatura en la base de datos y las categoría de cada una de ellas (indicando entre paréntesis la codificación utilizada en la base).

Variables clínico-epidemiológicas.

- Sexo: Sexo del paciente. Hombre(1) o Mujer(2).
- Edad: Edad del paciente al diagnóstico del melanoma.
- IMC: Índice de Masa Corporal del paciente al diagnóstico del melanoma en kg/m^2 .
- Fototipo: Clasificación según el tipo de piel Fitzpatrick, la cual clasifica las pieles según la reacción a la exposición solar. Dicha clasificación divide a la población en siete subgrupos:
 - 0: Hace referencia a los albinos.
 - I: Población de piel muy pálida, siempre se suelen quemar y nunca se broncean.
 - II: Población de piel clara, se suelen quemar con facilidad, pero tras mucha exposición solar suelen adquirir un discreto tono bronceado.
 - III: Población de piel morena clara, se suelen quemar en ciertas ocasiones, pero suelen adquirir un bronceado medio.

- IV: Población de piel morena, no se queman nunca y adquieren un tono bronceado intenso.
- V: Población de piel oscura, nunca se queman y el bronceado es muy intenso.
- VI: Población de piel negra, nunca se queman.

Aunque la clasificación incluya siete subgrupos, los datos que se manejan en este Trabajo de Fin de Máster provienen de la consulta del IVO donde no recibieron ningún paciente de fototipo 0 ni VI por lo que tan solo se analizarán los fototipos I-V.

- Ojos: Color de ojos del paciente. Oscuros(1) o Claros(2).
- Pelo: Color del pelo del paciente. Oscuro(1), Rubio(2) o Pelirrojo(3).
- Quemaduras: Historial de quemaduras intensas que hayan provocado ampollas o dolor durante al menos 48 horas. No(1) o Si(2).
- Añosprofsol: Años expuesto a la radiación solar debido a profesiones al aire libre.
- Añospaquete: Tabaquismo del paciente en años/paquete, unidad que se define como el número de años que habría fumado el paciente si hubiera fumado un paquete diario.
- Efélides: Presencia de pecas que aparecen en las zonas expuestas al sol. No(1) o Si(2).
- Lentigos: Presencia de hiperpigmentación o mancha en la piel producida por la exposición solar. No(1) o Si(2).
- Qactínicas: Presencia de queratosis actínicas, manchas asperas y escamosas en la piel. No(1) o Si(2).
- Segtumor: Desarrollo de un segundo tumor. No(1) o Si(2).
- CEC: Antecedentes personales de carcinoma epidermoide cutáneo. No(1) o Si(2).
- CBC: Antecedentes personales de carcinoma basocelular cutáneo. No(1) o Si(2).
- Angiomas_sen: Recuento de angiomas seniles (puntos rojos que aparecen en la piel). No(1), <10(2), 10-20(3), 21-50(4), 51-100(5) o >100(6).

- Queratosis_seborreicas: Recuento de queratosis seborreicas (lesiones de aspecto ceroso, escamoso y ligeramente elevado). No(1), <10(2), 10-20(3), 21-50(4), 51-100(5) o >100(6).
- Nevus: Recuento de nevus melanocíticos comunes (lunares). <20(1), 20-50(2), 51-100(3) o >100(4).
- Nevus atípicos: Presencia de nevus melanocíticos más grande de lo común. No(1) o Si(2).
- Múltiples melanomas: Presencia de múltiples melanomas en el paciente. No(1) o Si(2).
- Melanoma familiar: Antecedentes de melanoma en la familia. No(1) o Si(2).
- Cáncer familiar: Antecedentes de cualquier tipo de cáncer en la familia (ya sea de páncreas o cualquier otro). No(1) o Si(2).
- MC1Rvar: Número de variantes en el gen receptor de la melanocortina (MC1R). 0(0), 1(1) o >1(2).
- MC1R_RHC: Número de variantes 'R' en el gen MC1R. 0(0) , 1(1) o >1(2).
- R163Q. Presencia del polimorfismo asociado al receptor 1 de la melanocortina y la raza (del inglés *Race and melanocortin 1 receptor*). No(1) o Si(2).

Variables clínico-patológicas de caracterización del melanoma.

- Quemaduras_areaMM: Antecedentes de quemaduras en el área en la que se desarrolló el melanoma. No(1), Leves-Moderadas(2) o Graves(3).
- Lentigos_en_área_de_MM: Presencia de algún lentigo en el área del melanoma. No(1) o Si(2).
- Localización: Ubicación del melanoma. Cabeza/Cuello(1), Extremidad superior(2), Tronco(3), Extremidad inferior(4), Acral(5), Mucosos(6) o Primario desconocido(7).
- Fotolocalización: Localización del melanoma en función del patrón de exposición solar del área anatómica donde se encuentra el melanoma. Rara vez(1), Ocasional(2) o Habitual(3).

- TipoHx: Clasificación que describe como de rápido se podrían multiplicar las células cancerosas del tumor. Melanoma sobre lentigo maligno *LMM*(1), Melanoma de extensión superficial *MES*(2), Melanoma nodular *MN*(3), Melanoma lentiginoso acral *MLA*(4) o Otros/sin clasificar(5).
- Nevuspre: Restos de nevus en el melanoma. No(1) o Si(2).
- Elastosis: Signos de envejecimiento o de degradación en la dermis circundante al melanoma por la exposición solar acumulada. No(1), Ligero(2), Moderado(3) o Intenso(4).
- CSD: Reclasificación de la Elastosis según la clasificación de la OMS. Melanomas que se manifiestan en piel dañada cronicamente por el sol. No(1) o Si(2).

Mutaciones somáticas

- BRAF: Estado del gen BRAF, este produce una proteína que afecta a la diseminación y el crecimiento de las células. *wt*(0) o Mutado(1).
- NRAS: Estado del gen NRAS. Dicho oncogen produce proteínas que juegan una función muy importante en la división y la diferenciación celular. *wt*(0) o Mutado(1).
- KIT: Estado del gen KITT, este genera ciertas proteínas llamadas receptores de la tirosina quinasa, que controlan el crecimiento de las células y su división. *wt*(0) o Mutado(1).
- TERTprom: Estado del promotor del gen TERT. Este proporciona instrucciones para crear una encima llamada telomerasa. *wt*(0) o Mutado(1).

Características histológicas pronósticas.

- Breslow: Grado de invasión del tumor en la dermis (en mm).
- Ulceración: Ruptura en la piel que se encuentra sobre el melanoma. No(1) o Si(2).
- TIL: Cantidad de células de linfocitos infiltrantes en el tumor (del inglés, *tumor infiltrating lymphocytes*). Ausentes(1), Escasos(2) o Abundantes(3).
- Satelitosis: Presencia de acumulación de células inflamatorias alrededor de las células dañadas. No(1) o Si(2).

- Mitosis: Número de mitosis por milímetro cuadrado.
- Regresión: Ausencia de tumor en una zona donde se presentó el melanoma y que es sustituido por fibrosis. No(1) o Si(2).
- Invasión_vascular: Presencia de células de melanoma en el interior de un vaso. No(1) o Si(2).
- Ganglio_centinela: Primer ganglio de los linfáticos regionales que recibe de forma directa el drenaje linfático desde la piel. Negativo(1), Positivo(2) o No se identifica(3).
- Total_ganglios_positivos: Número de ganglios en los que se encontraron cáncer.

Estadio In situ(0), Localizada(1), Locorregional(2) o A distancia(3).

A.2. Marco teórico PCA

Aunque en la sección 3.2.2 se incluyese un breve resumen sobre el PCA, en el presente apartado se incluye una explicación matemática para el completo entendimiento del método.

Explicación matemática

Dada una base de datos, representada por la matriz $\mathbf{X} \in \mathcal{M}_{n \times k}(\mathbb{R})$, donde n representa el número de observaciones y k el número de variables, se desea transformar la matriz \mathbf{X} en otra matriz $\mathbf{T} \in \mathcal{M}_{n \times k}(\mathbb{R})$ mediante alguna matriz $\mathbf{P} \in \mathcal{M}_{n \times n}(\mathbb{R})$ tal que:

$$\mathbf{T} = \mathbf{P}\mathbf{X} \tag{A.1}$$

La ecuación (A.1) representa un cambio de base donde, \mathbf{X} está siendo proyectada sobre las columnas de \mathbf{P} ; por tanto, las filas de \mathbf{P} son una nueva base para representar las columnas de \mathbf{X} . Así pues, eligiendo una matriz \mathbf{P} adecuada, sus filas contendrán las direcciones de las componentes principales.

Como anteriormente ha sido explicado, uno de los supuestos del PCA es que las variables transformadas, las componentes principales, están incorrelacionadas. Lo que matemáticamente equivaldría a decir que la covarianza entre las variables debería de estar lo más cercana a 0 posible. Sin embargo, interesa que la varianza de las variables extraídas sea

lo mayor posible. Para conseguir este objetivo, se define \mathbf{S} como la matriz de varianzas-covarianzas de la matriz \mathbf{X} :

$$\mathbf{S} = \frac{1}{n-1} \mathbf{XX}^T \quad (\text{A.2})$$

Recuperando la ecuación (A.1), la matriz de varianzas-covarianzas de la matriz transformada \mathbf{T} se define como:

$$\mathbf{S}_T = \frac{1}{n-1} \mathbf{TT}^T = \frac{1}{n-1} (\mathbf{PX})(\mathbf{X}^T \mathbf{P}^T) = \frac{1}{n-1} \mathbf{PSP}^T \quad (\text{A.3})$$

Utilizando el teorema espectral del álgebra lineal, para una matriz \mathbf{A} real simétrica de dimensiones $n \times n$, sus n valores propios (denotados por λ_j) son reales y existe una matriz ortogonal real \mathbf{M} de dimensiones $n \times n$ que cumple:

$$\mathbf{A} = \mathbf{MDM}^{-1} \quad (\text{A.4})$$

Para una matriz diagonal \mathbf{D} cuyos valores en la diagonal corresponden a los valores propios de \mathbf{A} .

Dado que la matriz \mathbf{M} es ortogonal $\mathbf{Q}^{-1} = \mathbf{Q}^T$, si diagonalizásemos la matriz \mathbf{S} de la ecuación (A.3) se tendría:

$$\mathbf{S}_T = \mathbf{PSP}^T = \mathbf{P}(\mathbf{QDQ}^T)\mathbf{P}^T \quad (\text{A.5})$$

Por lo que si tomásemos $\mathbf{P} = \mathbf{Q}^T$ la matriz ortonormal, donde los vectores propios de \mathbf{S} estuvieran en las filas de \mathbf{P} :

$$\mathbf{S}_T = \mathbf{P}(\mathbf{QDQ}^T)\mathbf{P}^T = \mathbf{P}(\mathbf{P}^T \mathbf{D} \mathbf{P})\mathbf{P}^T = \mathbf{D} \quad (\text{A.6})$$

Se obtiene por tanto la matriz \mathbf{D} diagonal que tiene como valores los n valores propios (varianza explicada) correspondientes a los vectores propios (ejes principales) colocados en la matriz \mathbf{P}^T . Así pues, se consigue el objetivo anteriormente expuesto. Cabe mencionar que la matriz \mathbf{T} contiene a lo que llamaremos *Scores* (las proyecciones de las observaciones sobre las componentes) y que la matriz \mathbf{P} contiene a lo que haremos referencia como *Loadings*.

La obtención de las componentes principales puede también llevarse a cabo mediante la descomposición de valores singulares (*SVD*), pero este proceso no será explicado en este trabajo. Tampoco será introducido el algoritmo NIPALS, algoritmo secuencial para calcular las componentes principales, aunque en cierto punto pueda utilizarse.

Introducción de datos categóricos

Como anteriormente se ha explicado, el PCA es una técnica en la cual solo se permite el uso de variables continuas. Las bases de datos médicas suelen contener muchas variables categóricas por lo que habrá que realizar una pequeña transformación de los datos para poder incluirlas. Dicha transformación se basa en la creación de tantas variables *dummies* como categorías tenga la variable original de manera que recoja en ellas su información.

Se define como variable dummy a aquellas variables binarias que toman valor 0, 1 en función de pertenecer o no pertenecer a una categoría. Ejemplo: La variable *Sexo* es categórica se crean por tanto dos variables *dummies* para sustituirla, *Hombre* y *Mujer* definidas como:

$$Hombre = \begin{cases} 1 & \text{si } Hombre \\ 0 & \text{si } Mujer \end{cases} \quad \text{y} \quad Mujer = \begin{cases} 1 & \text{si } Mujer \\ 0 & \text{si } Hombre \end{cases}$$

Resultados

B.1. Validación de las imputaciones

En la Figura B.1, se puede apreciar como las dos opciones para las imputaciones realizadas son prácticamente idénticas. Además, si estas se comparan con los *Loadings* de la base de datos sin imputar y con la del algoritmo NIPLAS, más allá de seguir viendo el efecto espejo anteriormente mencionado, no se encontraron diferencias que hicieran saltar la alarma sobre una posible imputación de los datos errónea. Es más, analizando los gráficos de la R^2 explicada por cada variable en cada componente, no se avistan grandes diferencias en cuanto a la composición de las componentes se refiere. Por tanto, se concluye que las imputaciones realizadas son contundentes y que no hay evidencias de que podrían alterar los resultados que se obtendrían en caso de tener la base de datos completa.

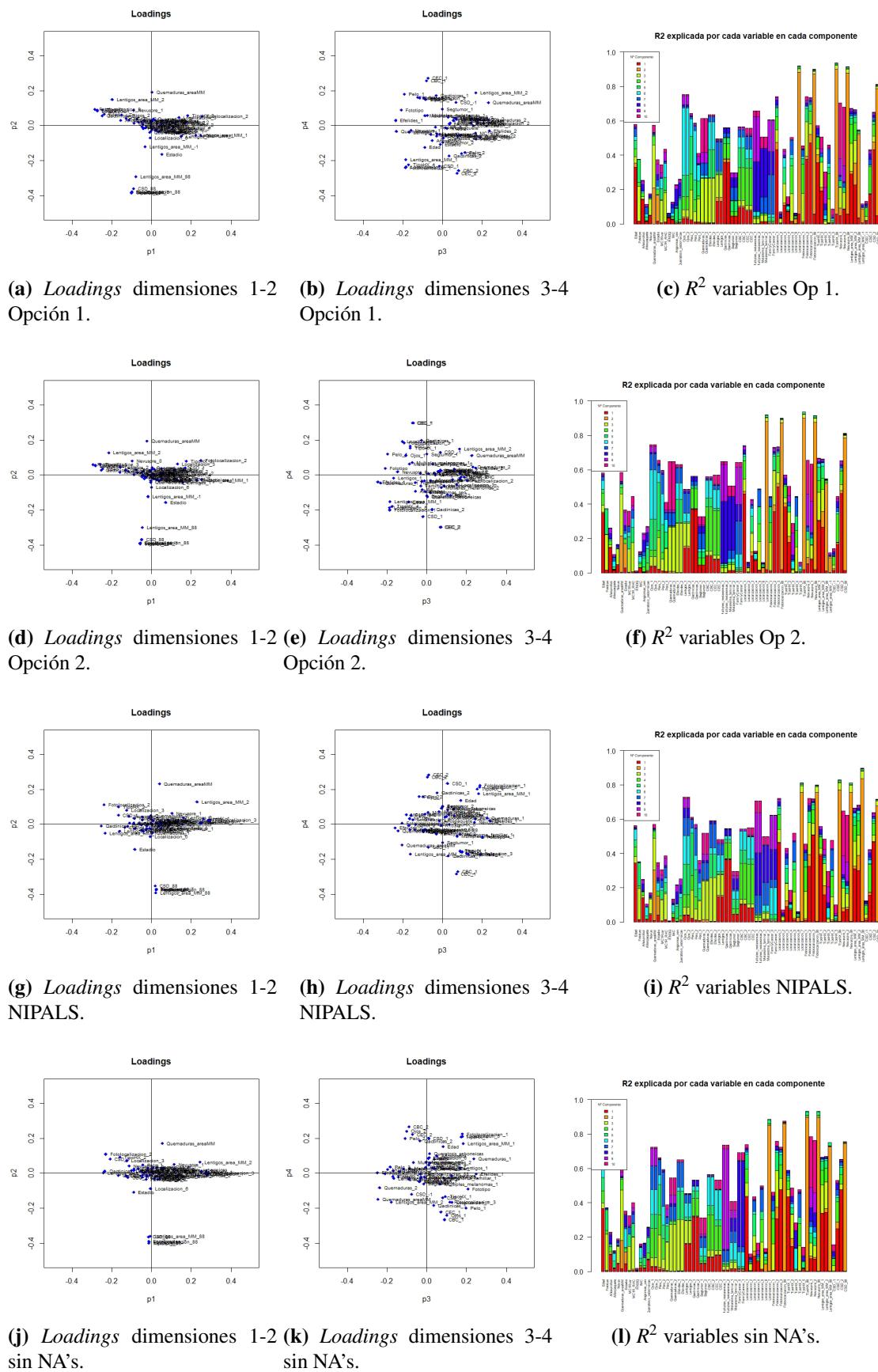


Figura B.1: Loading plots y R^2 explicada por cada variable en cada componente para las distintas bases de datos.

B.2. Clusters

B.2.1. Elección de número de clusters

A modo ilustrativo, se añaden los gráficos de los cuales se concluyó el número de clusters a utilizar en cada caso.

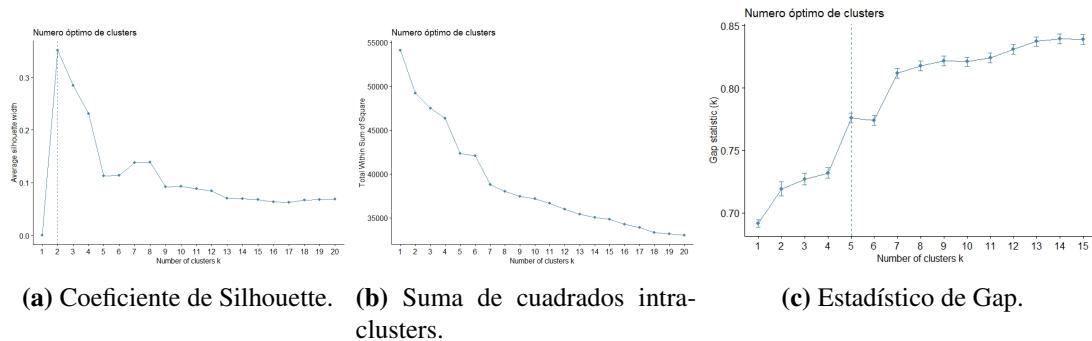


Figura B.2: Métodos de selección de número de *clusters* para distancia de Manhattan método *K-means complete*

En la Figura B.2, el índice de Silhouette (Figura B.2 (a)) indicó que lo adecuado sería la extracción de 2 *clusters*. Sin embargo, no parecía concordar con el *heatmap* (recordar Figura 4.8(b)). Además, el valor de la suma intra-clusters (Figura B.2 (b)) era muy elevado para esa opción, no parecía buena idea. Finalmente, el estadístico de Gap (Figura B.2 (c)) indicaba que deberían extraerse 5 clusters; sin embargo, en el Silhouette se presentaba un mínimo local para ese valor. Observando el siguiente valor para el cual se obtenía un cambio respecto al anterior, 7 *clusters* resultó la opción a considerar. En el resto de gráficas también parecía una buena opción; por lo que finalmente, se trajeron 7 *clusters* para este método.

Esta metodología puede ser utilizada tanto para la distancia euclídea como para la de Manhattan.

Para la elección del número óptimo de *clusters* en el caso de los métodos jerárquicos de la distancia Gower la metodología fue un tanto diferente. En la Figura B.3 se incluye un ejemplo.

Observando la Figura B.3, la extracción de 4 *clusters* parecía ser una opción correcta. Tanto el coeficiente de Silhouette (Figura B.3 (a)) como el índice de Dunn (Figura B.3 (b)) presentaron un máximo para dicho número. Además, para el índice de Frey (Figura

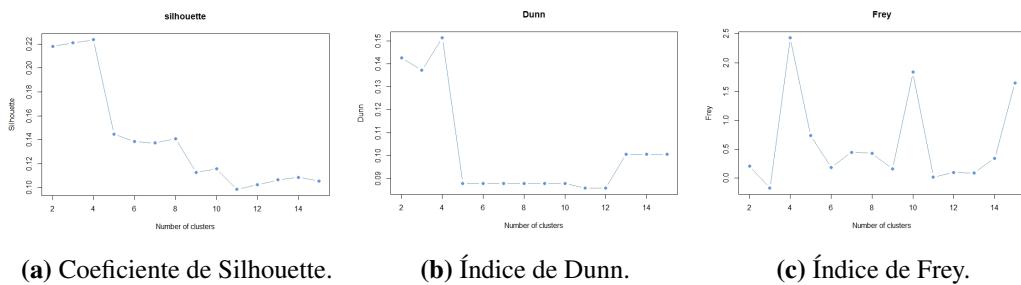


Figura B.3: Métodos de selección de número de *clusters* para distancia de Gower método *K-means Ward*.

B.3 (c)), se observó como era el valor anterior antes de que cayese a un valor inferior a 1. En consecuencia, se optó por la extracción de 4 *clusters*.

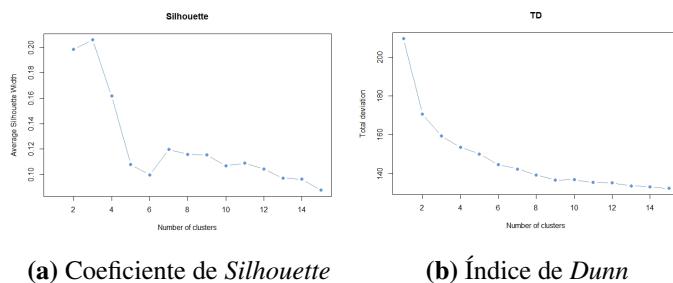


Figura B.4: Métodos de selección de número de *clusters* óptimo para distancia de Gower método *K-medoids*.

En el caso del método de *K-medoids* para la distancia de Gower, el coeficiente de Silhouette (Figura B.4 (a)) apuntó que la extracción de *cluster* 3 sería lo idóneo. Además, en la desviación total (Figura B.4 (b)), parecía crearse una especie de codo para ese mismo valor y los valores tampoco eran extremadamente altos por lo que no pareció mala opción.

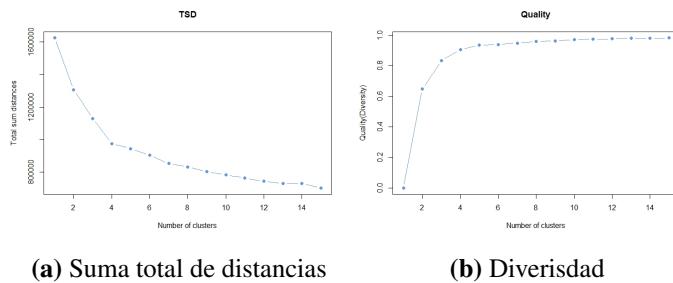


Figura B.5: Métodos de selección de número de clusters óptimo para método *K-prototypes*

Finalmente, se incluye la metodología que se utilizó para el método *K-prototypes*. En la Figura B.5 , en las gráfica suma total de distancias (Figura B.5 (a)) se creó una especie de codo para 4 *clusters*. Además, el gráfico de *Diversity* (Figura B.5 (c)), presentó un buen valor para 4 *clusters* que la extracción de más *clusters* no parecían mejorar. Finalmente, haciendo uso de la función *validation_kproto* incluida en *R* para el análisis del coeficiente de Silhouette para este método en concreto, la función indicó que el número óptimo de *clusters* a extraer era 4. Por tanto, fueron esos los clusters extraídos en este caso.

B.2.2. Clusters obtenidos para los grupos etiopatogénicos bien definidos

En la Tabla B.1, se presentan los *clusters* creados mediante el método *K-means* para la distancia de Manhattan.

Tabla B.1: Distribución del grupo etiopatogénico en los *clusters* para el método *K-means* mediante distancia de Manhattan.

	Acral	CSD	Mixto	Mucoso	Nevogénico	Nevogénico débil	Primario desconocido
Cluster 1	0	224	36	0	2	8	0
Cluster 2	88	1	1	15	9	15	0
Cluster 3	1	3	6	3	179	195	1
Cluster 4	0	0	0	1	0	0	33

Claramente se pudo observar como el método aislabía a los *Primario desconocidos*, mientras que creaba agrupaciones del resto de grupos etiopatogénicos. Mezclando así los *Mixtos* y *CSD* por un lado, los *Nevogénicos* y los *Nevogénico débil* por otro y los *Acrales* y los *Mucosos* finalmente.

De igual manera, en la Tabla B.2 se observa la agrupación que creó la metodología de *Fuzzy K-means*.

Tabla B.2: Distribución del grupo etiopatogénico en los *clusters* para el método de *Fuzzy K-means*.

	Acral	CSD	Mixto	Mucoso	Nevogénico	Nevogénico débil	Primario desconocido
Cluster 1	0	7	10	2	182	202	1
Cluster 2	89	1	1	16	5	9	0
Cluster 3	0	220	32	0	3	7	0
Cluster 4	0	0	0	1	0	0	33

En este caso también, se apreció la misma tendencia a agrupación de los individuos encontrada en la Tabla B.1. Por tanto, se observó cierta consistencia en los resultados.

Por último, en la Tabla B.3 se presentan los resultados del cluster jerárquico mediante el método de Ward para la distancia de Manhattan.

Tabla B.3: Distribución del grupo etiopatogénico en los *clusters* para el método jerárquico mediante distancia de Manhattan.

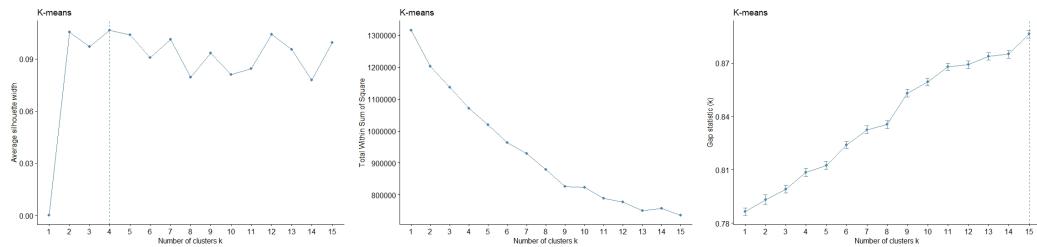
	Acral	CSD	Mixto	Mucoso	Nevogénico	Nevogénico débil	Primario desconocido
Cluster 1	0	205	32	0	3	7	0
Cluster 2	2	9	7	1	171	193	1
Cluster 3	86	0	0	0	3	4	0
Cluster 4	0	0	0	0	0	0	33
Cluster 5	1	3	3	1	13	11	0
Cluster 6	0	11	1	0	0	3	0
Cluster 7	0	0	0	17	0	0	0

Este caso es un tanto más embrolloso. Los *Mucosos*, *Acrales* y *Primario desconocido* se aislaron cada uno en un *cluster* diferente. Además, los *CSD* y *Mixto* por un lado y los *Nevogénico* y *Nevogénico débil* por otro seguían juntándose. Sin embargo, se crearon dos *clusters* adicionales en los que se encontraron individuos de diferentes procedencias. Tenían cierto aire a observaciones anómalas.

B.3. Obtención de clusters para grupos 'indefinidos'

Como se hizo con los grupos etiopatogénicos bien definidos, para la creación de los *clusters* mediante las distintas técnicas fue necesario definir primero el número óptimo de *clusters* a extraer por cada método. A continuación, se añade de manera ilustrativa el proceso para la distancia de Manhattan mediante *K-means* puesto que fue este el que creo conflicto a la hora de la elección del número de *clusters* a extraer.

Como se observa en la Figura B.6, los diferentes índices no concluyeron claramente el número de clusters a extraer. Mientras que el coeficiente de Silhouette (Figura B.6 (a)) parecía indicar 4 *clusters*, la suma de cuadrados intra-cluster (Figura B.6 (b)) parecía ser muy elevada. Otra opción, parecía la extracción de 9 *clusters*, valor para el cual se observó una especie de codo para la suma de cuadrados intra-cluster.



(a) Coeficiente de Silhouette. **(b)** Suma de cuadrados intra-clusters. **(c)** Estadístico de Gap.

Figura B.6: Métodos de selección de número de *clusters* para la distancia de Manhattan mediante el método *K-means* para pacientes de grupos etiopatogénicos difusos.

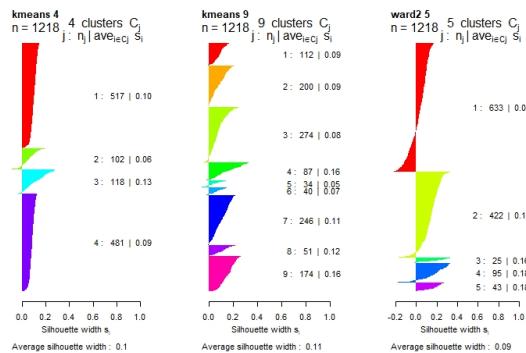


Figura B.7: Coeficiente de Silhouette para las diferentes técnicas para pacientes de grupos 'indefinidos'.

Como añadido y con tal de tener una mejor visión de los resultados, analizando el coeficiente de Silhouette (Figura B.7), el método de *K-means* para 9 *clusters* presentó el mayor valor con un 0.11; sin embargo, para 4 *clusters* el valor no fue muy inferior (0.1) y teniendo en cuenta el heatmap presente en la sección 4.4.2 cobraba más sentido optar por este último. Dada la naturaleza de los algoritmos, no se pudo añadir la gráfica para el método *Fuzzy K-means*; no obstante, se pudo calcular que su valor aumentaba hasta un 0.15. Este valor también se tuvo en cuenta a la hora de tomar la decisión de decantarse por esta metodología como se presentó en la sección 4.4.2.

Finalmente, en la Tabla B.4, se añade la cantidad de pacientes *No clasifiable* y *Non-risky* en los tres grupos extraídos mediante el método *Fuzzy K-means*.

Como se puede observar en dicha tabla, ninguno de los *clusters* creados pareció contener mayoritariamente ninguno de los grupos etiopatogénicos de referencia.

Tabla B.4: Pacientes según *Grupo etiopatogénico* y grupos creados

	No clasificable	Non-risky
Pseudo Non risky	142	305
Pseudo CSD	130	201
Pseudo Nevogénico	257	183

B.3.1. Comprobación resultados

Para comprobar los resultados obtenidos en la sección 4.4.3, en esta sección se incluyen los resultados de los estudios univariados llevados a cabo.

Cabe recordar que como ya se explicó en secciones anteriores, para calcular el test de Fisher aplicado en este caso, se excluyeron del estudio los pacientes con valores no aplicables (n.a.) y desconocido por no ser relevantes para nuestro estudio. Sin embargo, a la hora de presentar los resultados, si que se dejarán las proporciones de los valores n.a.

Analizando los resultados de la Tabla B.5, se pudo observar como los resultados concordaron del todo con los resultados expuestos en la sección 4.4.3. Como añadido, cabe comentar los resultados de la *Elastosis*, variable que no había sido estudiada hasta el momento. Dicha variable, presentó diferencias estadísticamente significativas. Analizando las proporciones de las categorías, se puede observar como el *Pseudo CSD* presenta valores más elevados que el resto de los grupos, siendo el *Pseudo Non-risky* el que menor *Elastosis* presenta entre los tres grupos.

Tabla B.5: Distribuciones de las variables en los grupos creados

Variables	Categorías	Tamaño muestral (n = 1218)			p-valor	p ajustado
		Pseudo Non Risky	Pseudo CSD	Pseudo Nevogénico		
Edad	0-30	0.1298	0.0121	0.0682	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	31-65	0.6443	0.3353	0.7614		
	>66	0.226	0.6526	0.1705		
Fototipo	I	0.0089	0.0332	0.0682	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	II	0.1857	0.3112	0.4045		
	III	0.3535	0.3384	0.3568		
	IV	0.4228	0.3021	0.1659		
	V	0.0291	0.0151	0.0045		
Añosprofsol	0	0.8479	0.6133	0.825	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	0-30	0.1141	0.145	0.1409		
	>30	0.038	0.2417	0.0341		
Añospaquete	0	0.613	0.5257	0.4659	1.89e ⁻¹¹	6.05e ⁻¹⁰
	0-30	0.3043	0.2417	0.4182		
	30-60	0.0537	0.145	0.0795		
	60-80	0.0134	0.0665	0.0205		
	>80	0.0157	0.0211	0.0159		
Ojos	Oscuros	0.7763	0.4834	0.5636	<2.2e ⁻¹⁶	2.9e ⁻¹⁶
	Claros	0.2237	0.5166	0.4364		
Pelo	Oscuro	0.9195	0.7553	0.6886	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Rubio	0.0783	0.2024	0.2068		
	Pelirrojo	0.0022	0.0423	0.1045		
Quemaduras	No	0.7248	0.568	0.225	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.2752	0.432	0.775		
Efélides	No	0.8971	0.7674	0.4432	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.1029	0.2326	0.5568		
Lentigos	No	0.3937	0.003	0.0045	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.6063	0.997	0.9955		
Q. actínicas	No	0.9911	0.6073	0.95	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.0089	0.3927	0.05		
Segundo tum	No	0.9329	0.6888	0.9114	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.0671	0.3112	0.0886		
CBC	No	0.9732	0.8278	0.9523	5.66e ⁻¹⁵	1.81e ⁻¹³
	Sí	0.0268	0.1722	0.0477		
CEC	No	0.9978	0.9486	0.9841	7.67e ⁻⁶	0.0002
	Sí	0.0022	0.0514	0.0159		
Nevus	<20	0.9105	0.9275	0.7091	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	20-50	0.0492	0.0393	0.1432		
	51-100	0.0336	0.0211	0.0955		
	>100	0.0067	0.0121	0.0523		
Múltiples melanomas	No	0.9732	0.9366	0.925	0.0045	0.1451
	Sí	0.0268	0.0634	0.075		
Melanoma familiar	No	0.9396	0.9305	0.8795	0.0029	0.0938
	Sí	0.0604	0.0695	0.1205		
Cáncer familiar	No	0.4899	0.4808	0.5114	0.67	1
	Sí	0.5101	0.5196	0.4886		

Tabla B.6: Distribuciones de las variables en los grupos creados. Continuación Tabla B.5

Variables	Categorías	Tamaño muestral (n = 1218)			p-valor	p ajustado
		Pseudo Non Risky	Pseudo CSD	Pseudo Nevogénico		
Quemaduras área MM	No	0.6242	0.3263	0.1364	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Leves-Moderadas	0.3177	0.4592	0.4591		
	Graves	0.0582	0.2145	0.4045		
Localización	Cabeza/cuello	0.0694	0.3353	0.0205	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Tronco	0.1186	0.1813	0.2023		
	Extr inferior	0.3557	0.3746	0.5909		
	Acral	0.3624	0.0906	0.1795		
	Mucoso	0.094	0.0181	0.0068		
Tipo histológico	LLM	0.0179	0.2589	0.0182	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	MES	0.6689	0.4048	0.8705		
	MN	0.2036	0.2659	0.0818		
	Otros	0.1096	0.0695	0.0295		
Estadio	In situ	0.1051	0.2145	0.1932	0.0005	0.016
	Localizada	0.6756	0.6133	0.7136		
	Locorregional	0.2103	0.1631	0.0909		
	A distancia	0.0089	0.0091	0.0023		
MC1R	0	0.4474	0.3384	0.2364	6.44e ⁻¹⁵	2.06e ⁻¹³
	1	0.4273	0.4683	0.4364		
	>1	0.1253	0.1934	0.3273		
RHC	0	0.8702	0.7583	0.625	1.11e ⁻¹⁶	3.56e ⁻¹⁵
	1	0.1275	0.2175	0.3091		
	>1	0.0022	0.0242	0.0659		
R163Q	No	0.9732	0.9728	0.9544	0.2263	1
	Sí	0.0268	0.0272	0.0456		
Fotolocalización	Rara vez	0.1834	0.0876	0.0182	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Ocasional	0.7785	0.4864	0.9614		
	Habitual	0.038	0.426	0.0205		
Lentigos en el área del MM	No	0.9329	0.2447	0.2841	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.0515	0.719	0.6932		
	n.a. (n=29)	0.0157	0.0363	0.0227		
CSD	No	0.9812	0.9452	0.9947	2.86e ⁻⁵	0.0009
	Sí	0	0.0457	0.0053		
	n.a.(n=8)	0.0188	0.0091	0		
	v.f. (n= 491)					
Nevuspre	No	0.783	0.9003	0.6045	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Sí	0.217	0.0997	0.3955		
Angiomás seniles	1	0.5134	0.3022	0.4039	1.03e ⁻⁵	0.0023
	2	0.2582	0.3111	0.3127		
	>2	0.2285	0.3867	0.2834		
	v.f. (n=349)					
Queratosis seborreicas	1	0.6042	0.2978	0.557	4.45e ⁻¹³	1.42e ⁻¹¹
	2	0.2143	0.2533	0.2378		
	>2	0.1815	0.4489	0.2052		
	v.f. (n=350)					
log(IMC) ¹		3.2297	3.2905	3.2323	0.81	1
Elastosis	No	0.8215	0.4911	0.7835	<2.2e ⁻¹⁶	<2.2e ⁻¹⁶
	Ligero	0.1169	0.25	0.1804		
	Moderado	0.0369	0.1920	0.0361		
	Intenso	0	0.058	0		
	n.a.(n=10)	0.0246	0.0089	0		
v.f.(n=475)						

B.4. Supervivencia

B.4.1. Estudios univariados

A continuación, se añaden gráficas de curvas de Kaplan-Meier en las que se estudiaron la supervivencia global y específica al melanoma según ciertas variables de interés de manera univariada. En ellas, se puede encontrar el p-valor asociado al log-rank test realizado.

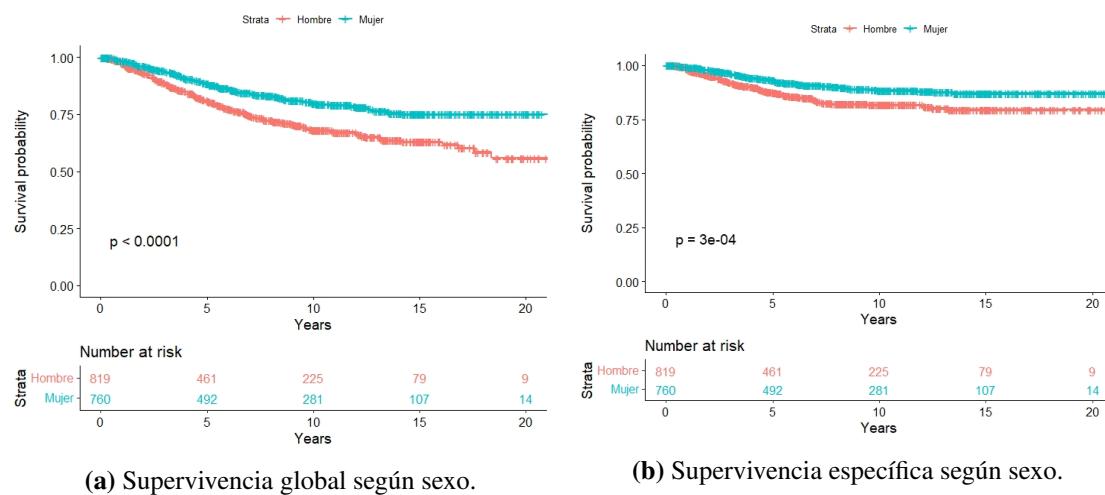


Figura B.8: Curvas de Kaplan-Meier para estudio de supervivencia según sexo.

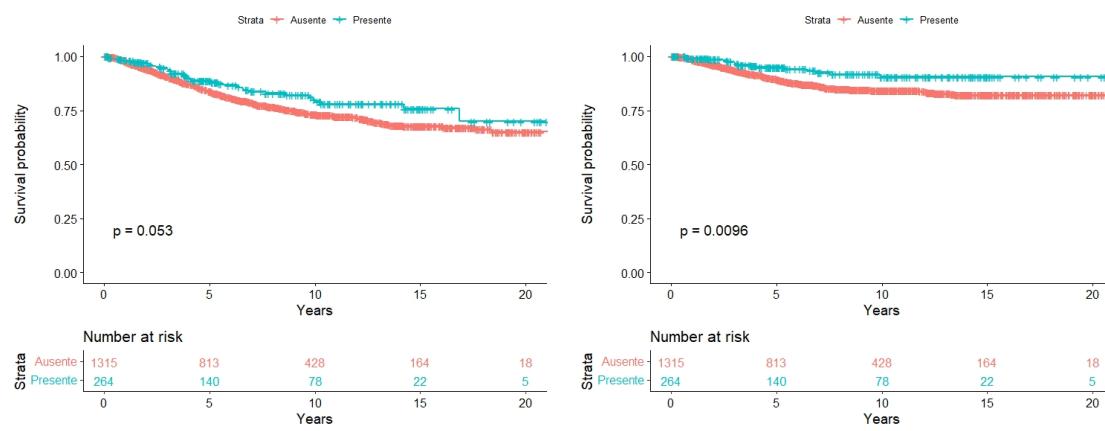
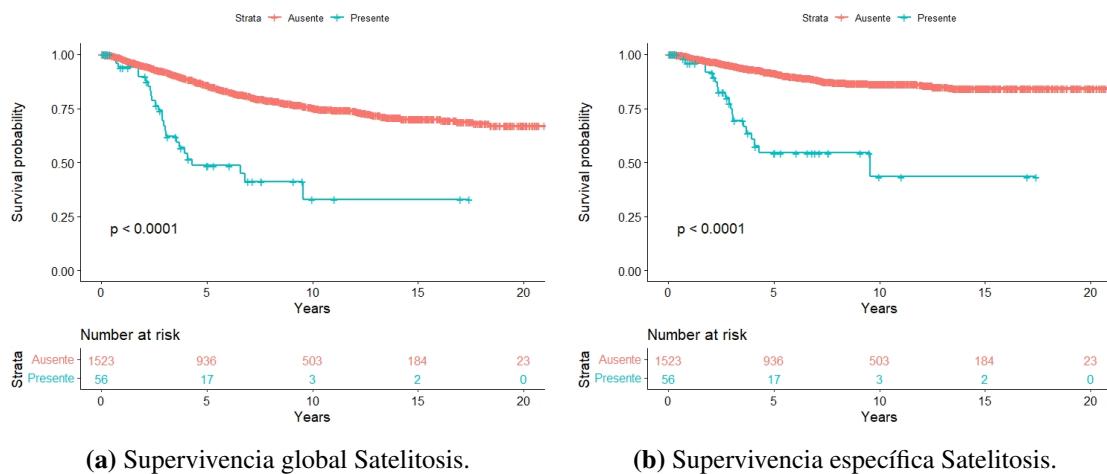


Figura B.9: Curvas de Kaplan Meier para estudio de supervivencia por regresión.

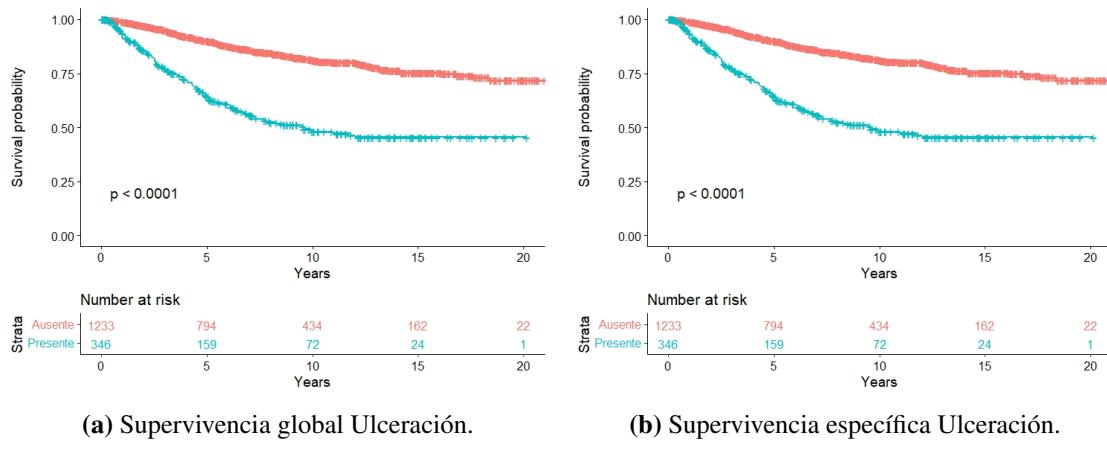
En la Figura B.8, se pudo observar como las mujeres presentaron una mejor supervivencia en los dos casos (tanto en la supervivencia global como en la específica). En cuanto a la regresión se refiere, analizando la Figura B.9, se observó como si fuesemos puristas

la variable no presentaba diferencias estadísticamente significativas en la supervivencia global; sin embargo, si que las presentó en la supervivencia específica. En este caso, la presencia de la regresión era la que ofrecía el efecto protector.



(a) Supervivencia global Satelitosis.

(b) Supervivencia específica Satelitosis.

Figura B.10: Curvas de Kaplan-Meier para estudio de supervivencia para satelitosis.

(a) Supervivencia global Ulceración.

(b) Supervivencia específica Ulceración.

Figura B.11: Curvas de Kaplan-Meier para estudio de supervivencia para ulceración.

En las Figura B.10 y B.11, se pudo observar como en ambos casos para los dos tipos de supervivencia se presentaron diferencias estadísticamente significativas siendo la ausencia tanto de *Satelitosis* como de *Ulceración* las que presentaban un efecto protector.

Como se pudo ver en la Figura B.12, en el caso de *TIL* no se encontraron diferencias estadísticamente significativas en ninguna de las supervivencias.

En el caso de la Invasión vascular (Figura B.13), al igual que en los casos de *Satelitosis* y *Ulceración*, se encontraron diferencias estadísticamente significativas siendo la ausencia de esta protectora frente a la muerte.

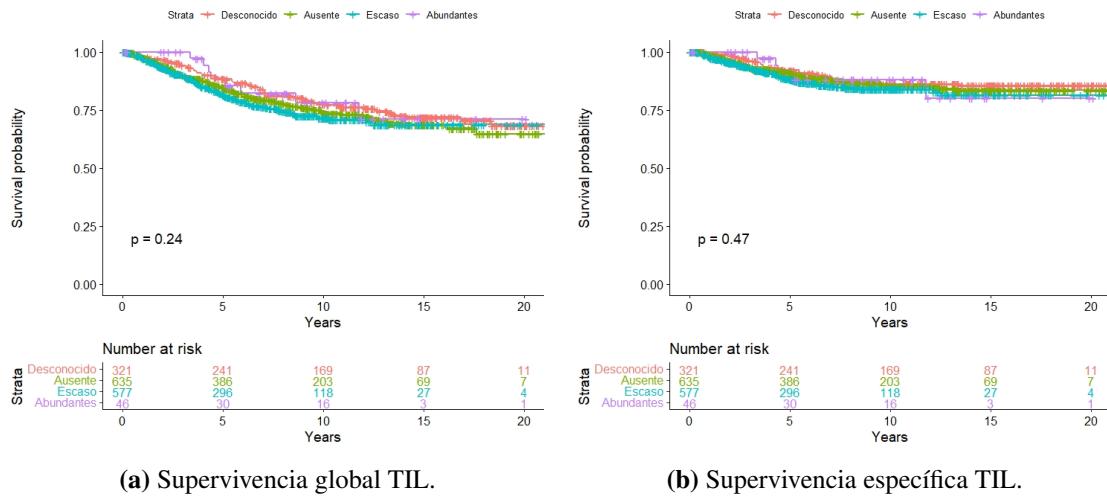


Figura B.12: Curvas de Kaplan Meier para estudio de supervivencia según TIL.

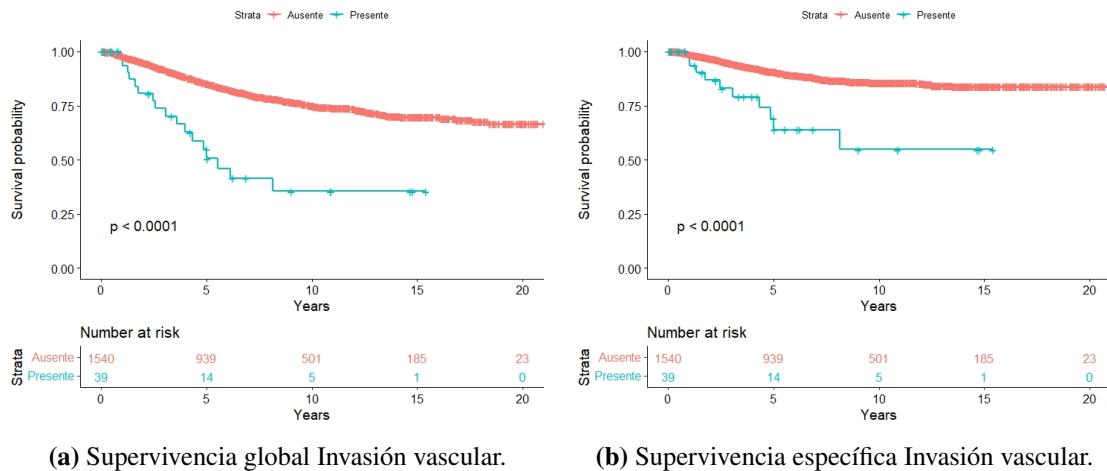


Figura B.13: Curvas de Kaplan-Meier para estudio de supervivencia para Invasión vascular.

Estudiando las figuras B.14 y B.15, se pudo ver como las diferencias estadísticamente significativas se volvieron a hacer notorias. Además, en ambos casos, valores mayores tanto de *Breslow* como de *Mitosis* presentaron una peor supervivencia que valores bajos o nulos de ellas.

Lo mismo ocurrió con el *Total de ganglios positivos* como se puede observar en la Figura B.16.

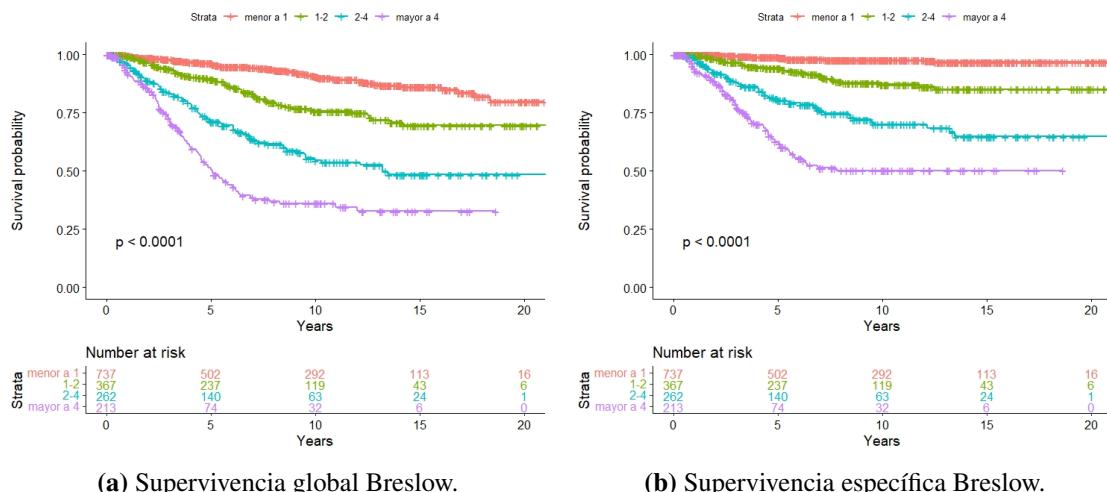


Figura B.14: Curvas de Kaplan-Meier para estudio de supervivencia según Breslow.

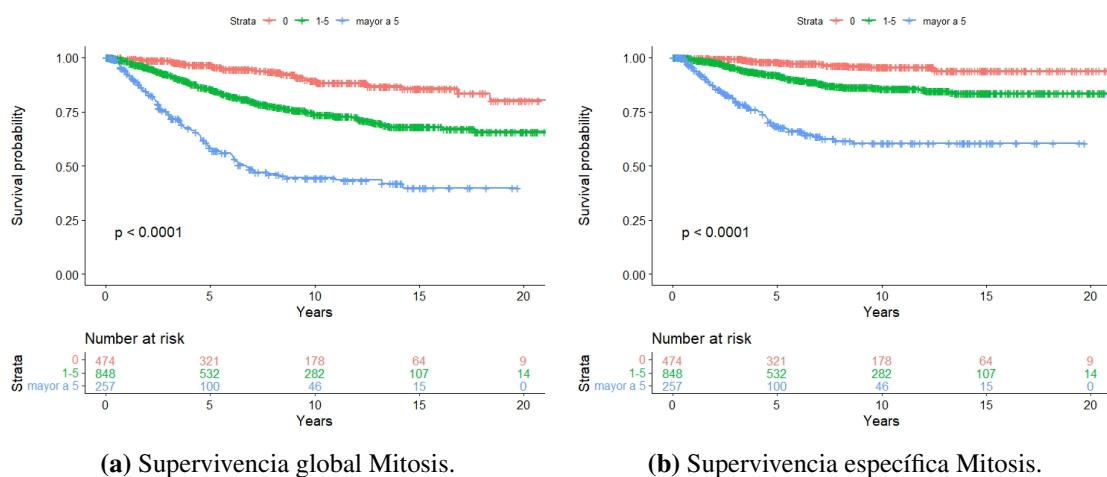
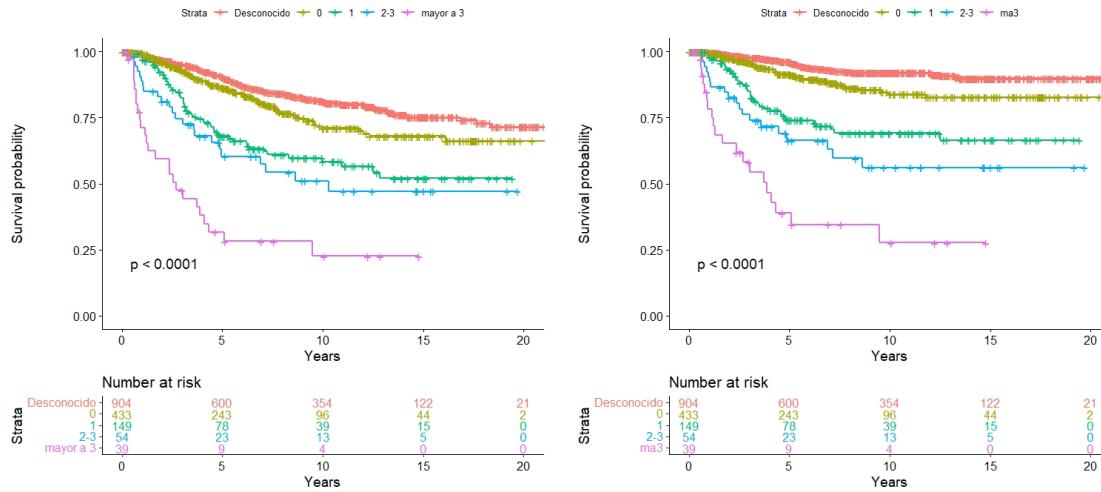


Figura B.15: Curvas de Kaplan-Meier para estudio de supervivencia para Mitosis.

B.4.2. Estudio multivariado

Para apoyar la validación de los modelos multivariados presentados en las secciones 4.6.2 y 4.6.3, se añaden los estudios de validación realizados sobre los modelos finales tanto de la supervivencia global (Tabla B.7 y Figura B.17) como de la supervivencia específica (Tabla B.8 y Figura B.18). Estos primeros resultados son los de los modelos que mantenían los grupos *Pseudo Nevogénico* y *Pseudo CSD* separados de los grupos etiopatogénicos a los que se asemejaban.

Como se puede observar en la Tabla B.7, se cumplieron los supuestos de proporcionalidad de los *hazards* para las variables incluidas en el modelo. Además, observando la Figura B.17, se vio como en este caso tan solo 41 individuos superaron el límite de 2, por simple



(a) Supervivencia global Total de ganglios positivos.

(b) Supervivencia específica Total de ganglios positivos.

Figura B.16: Curvas de Kaplan-Meier para estudio de supervivencia según Total de ganglios positivos.

Tabla B.7: Residuos Schoenfeld para el modelo multivariante de supervivencia global.

	p-valor
Etiogrups	0.691
Breslowcut	0.893
Mitosis d.t.	0.508
Global	0.85

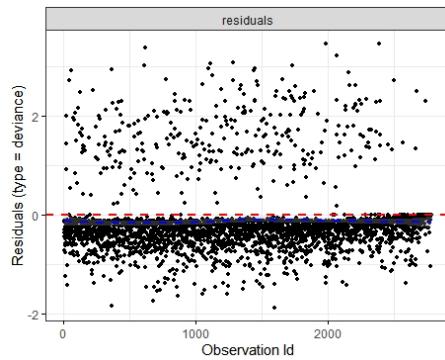


Figura B.17: Residuos deviance para el modelo multivariante de supervivencia global.

azar se esperaría que 78 pacientes lo superasen por lo que no hay de que preocuparse. Del mismo modo, en la supervivencia específica, en la Tabla B.8 se observa como se vuelve a cumplir la proporcionalidad de los *hazards* y una vez más, la cantidad de pacientes que despuntaron en la Figura B.18 no era de preocuparse.

En la Tabla B.9 se puede ver como el modelo final para la supervivencia global en caso de juntar los grupos *Pseudo Nevogénico* con los grupos etiopatogénicos *Nevogénico* y *Nevogénico débil* por una parte y los *Pseudo CSD* con los grupos etiopatogénicos *CSD* y *Mixto* por otro, cumplían con la proporcionalidad de los *hazards*. Además, en la Figura B.19 no se observaron pacientes que despuntasen y que podrían resultar un problema para

Tabla B.8: Residuos *Schoenfeld* para el modelo multivariante de supervivencia específica.

	p-valor
Breslowcut	0.93
Ulceración.	0.23
Global	0.6

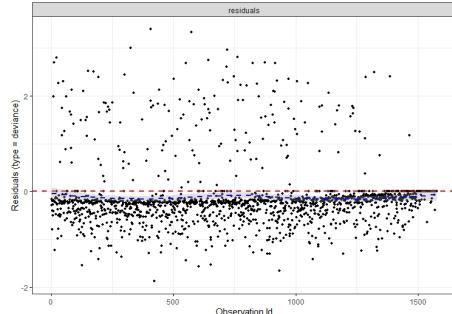


Figura B.18: Residuos deviance para el modelo multivariante de supervivencia específica.

Tabla B.9: Residuos *Schoenfeld* para el modelo multivariante de supervivencia global para los grupos etiopatogénicos juntados

	p-valor
Etiogrups	0.48
Breslowcut	0.80
Mitosis d.t.	0.36
Global	0.65

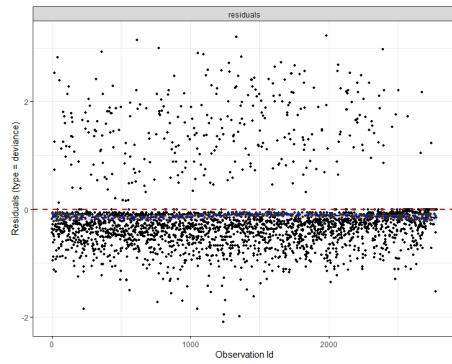


Figura B.19: Residuos deviance para el modelo multivariante de supervivencia global para los grupos etiopatogénicos juntados

el modelo. De hecho, fueron 75 los pacientes que superaron el valor de 2. Nada por lo que alarmarse.

ANEXO C

Funciones desarrolladas

C.1. Limpieza base de datos

C.1.1. *dummytovariable*

```
1 dummytovariable<-function(var1, var88){  
2     #var1: variable dummy principal  
3     #data: variable dummy con no calculables  
4     var88[intersect(which(is.na(var88)),which(!is.na(var1)))]<-0  
5     var<-NA  
6     levels(var)<-c(levels(var), '0')  
7     levels(var)<-c(levels(var), '1')  
8     levels(var)<-c(levels(var), '88')  
9     for (i in 1:length(var1)){  
10         if (var1[i]==1){  
11             var[i]<-'1'  
12         }  
13         else{  
14             if(var88[i]==1){  
15                 var[i]<-'88'  
16             }  
17             else{  
18                 var[i]<-'0'  
19             }  
20         }  
21     }  
22     var<-as.factor(var)  
23     return(var)}
```

```

1 dummy2tovariable<-function(var2, var3, var88){
2   #var2: variable dummy con una de las categorías
3   #var3: variable dummy con otra de las categorías
4   #data: variable dummy con no calculable
5   var88[intersect(which(is.na(var88)),which(!is.na(var2)))]<-0
6
7   var<-NA
8   levels(var)<-c(levels(var), '1')
9   levels(var)<-c(levels(var), '2')
10  levels(var)<-c(levels(var), '3')
11  levels(var)<-c(levels(var), '88')
12  for (i in 1:length(var2)){
13    if (var2[i]==1){
14      var[i]<-'2'
15    }
16    else if(var3[i]==1) {
17      var[i]<-'3'
18    }
19    else if(var88[i]==1){
20      var[i]<-'88'
21    }
22    else{
23      var[i]<-'1'
24    }
25  }
26  var<-as.factor(var)
27  return(var)
28 }
```

C.2. Validación de las imputaciones

C.2.1. plotloading

```

1 plotloading<-function(pca, i, j){
2   #pca: objeto PCA de opls
3   #i: componente a dibujar en el eje X
4   #j: componente a dibujar en el eje Y
5   plot(pca@loadingMN[,i], pca@loadingMN[,j], main = "Loadings",
6   xlab = paste('p',i, sep=''), ylab = paste('p',j, sep=''),
7   pch = 18, col = "blue", xlim=c(-0.5,0.5) ,ylim=c(-0.5,0.5))
8   # Asignamos las etiquetas
9   text(pca@loadingMN[,i], pca@loadingMN[,j], labels = row.names(pca@loadingMN),
10  cex = 0.6, pos = 4, col = "black")
11  abline(h=0, v=0)
12 }
```

C.2.2. R2varcomp

```

1 R2varcomp<-function(pca, data){
2   X = as.matrix(data)
3   SCT=nrow(data)-1
4   mat<-NA
5   for (i in 1:pca@summaryDF$pre){
6     pca.o<-opls(data, scaleC='none', predI=i, info.txtC='none', fig.pdfC='none')
7     Loadings=pca.o@loadingMN
8     Scores = pca.o@scoreMN
9     Xest=Scores %*% t(Loadings)
10    SCE=colSums(Xest**2)
11    R2=SCE/SCT
12    mat<-rbind(mat, R2)
13  }
14  mat<-mat[-1,]
15  mat1<-mat
16  for (i in 2:dim(mat)[1]){
17    mat1[i, ]=mat[i,]-mat[i-1,]
18  }
19  rownames(mat1)<-c(1:pca@summaryDF$pre)
20  barplot(mat1, las=2, col=c(1:nrow(mat1)), ylim=c(0,1), main='R2 explicada por cada'
21  variable en cada componente', cex.names = 0.5)
22  legend(x = "topright", legend = rownames(mat1), fill = 1:nrow(mat1),
23  title = "N Componente", xpd = TRUE, cex=0.5)
24}

```

C.2.3. SCR

```

1 SCR<-function(pca, data){
2   #pca: modelo pca creado mediante rpls
3   #data: base de datos utilizada para la creación del modelo
4   X = as.matrix(data)
5   Loadings=pca@loadingMN
6   Scores = pca@scoreMN
7   E = X - Scores %*% t(Loadings)
8   Scr = rowSums(E^2)
9   plot(1:length(Scr), Scr, type = "l", ylab = "SCR", xlab = "patients")
10  g = var(Scr)/(2*mean(Scr))
11  h = (2*mean(Scr)^2)/var(Scr)
12  chi2lim = g*qchisq(0.95, df = h)
13  chi2limi2=g*qchisq(0.99, df=h)
14  abline(h = chi2lim, col = 2, lty = 2)
15  abline(h=chi2limi2, col=4, lty=2)
16  atípicas = which(Scr > chi2lim)
17  atípicas2 = which(Scr > 2*chi2lim)
18  d<-list(atípicas=atípicas,atípicas2=atípicas2, E=E, scr=Scr)
19  return(d)  }

```

C.2.4. T2

```

1   T2<-function(pca){
2     #pca: modelo pca creado mediante roplis
3     K = pca@summaryDF[[2]]
4     Scores = pca@scoreMN
5     T2 = colSums(t(Scores**2) / pca@pcaVarVn)
6     N = as.numeric(pca@descriptionMC[[1]])
7     F95 = K*(N**2 - 1)/(N*(N - K)) * qf(0.95, K, N-K)
8     F99 = K*(N**2 - 1)/(N*(N - K)) * qf(0.99, K, N-K)
9     plot(1:length(T2), T2, type = "l", xlab = "patients", ylab = "T2")
10    abline(h = F95, col = "orange", lty = 2, lwd = 2)
11    abline(h = F99, col = "red3", lty = 2, lwd = 2)
12    anomalas = which(T2 > F95)
13    anomalas2 = which(T2 > 2*F95)
14    d<-list(anomalas=anomalas,anomalas2=anomalas2)
15    return(d)
16
17  }

```

C.2.5. Contri

```

1   Contri<-function(i, E, SCR){
2     #i: n del individuo del que se quiere obtener la contribución a la SCR
3     #E: matriz de residuos del modelo, calculada por: E = X - Scores %*% t(Loadings)
4     #SCR: suma de los cuadrados de los errores, calculada por: SCR = rowSums(E^2)
5     contribucion<-NA
6
7     for (j in 1:length(SCR)){
8       eind<-E[j,]
9       signo<-sign(eind)
10      contri<-(signo*(eind^2)/SCR[j])*100
11      contribucion<-rbind(contribucion,contri)
12    }
13    contribucion<-contribucion[-1,]
14
15    if (is.null(rownames(E)[i]))
16    {
17      return(barplot(contribucion[i,],las=2, cex.names = 0.45, main=c('Contribuciones a SCR
18      individuo',i)))
19    }
20    else{
21      return(barplot(contribucion[i,],las=2, cex.names = 0.45, main=c('Contribuciones a SCR
22      individuo',rownames(E)[i])))
23    }
24  }

```

C.3. Clustering

C.3.1. hopkins

```

1 hopkins<-function(data,m, seed, method){
2   #data: takes the dataframe to be used
3   #m: number of points to take into account in the hopkins statistic
4   #seed for the random data observations to take into account
5   #method: distance to be used
6   set.seed(seed)
7   if (method=='gower'){
8     namx<-sample(rownames(data), m, replace = FALSE)
9     sub<-data[namx,]
10    gower_dist = gower.dist(sub)
11    dista<-as.matrix(as.dist(gower_dist))
12    diag(dista[,rowMins(dista, value=FALSE)])<-Inf
13    sx<-sum(rowMin(dista))
14    random_df<-as.data.frame(matrix(0, nrow=dim(data)[1], ncol = dim(data)[2]))
15    for (i in 1:dim(data)[2]){
16      if (class(data[,i])=='numeric'){
17        random<-as.data.frame(runif(length(data[,i]),min(data[,i]), max(data[,i])))
18        random_df[,i]<-random
19      }
20      else{
21        #generate random numbers with a discrete uniform distribution
22        random_df[,i]<-sample(levels(data[,i]),dim(data)[1], replace = TRUE)
23        random_df[,i]<-as.factor(random_df[,i])
24      }
25    }
26    namy<-sample(rownames(random_df), m, replace = FALSE)
27    suby<-random_df[namy,]
28    gower_disty = gower.dist(suby)
29    distay<-as.matrix(as.dist(gower_disty))
30    diag(distay[,rowMins(distay, value=FALSE)])<-Inf
31    sy<-sum(rowMin(distay))
32  }
33  else{
34    namx<-sample(rownames(data), m, replace = FALSE)
35    dista <- as.matrix(dist(data, method = method))
36    diag(dista)<-Inf
37    sx<-sum(rowMin(dista[namx,]))
38    suby <- as.data.frame(apply(data, 2, function(x, m){runif(m, min(x), max(x))}, m))
39    maty<-rbind(data, suby)
40    distay <- as.matrix(dist(maty, method = method))
41    diag(distay)<-Inf
42    sy<-sum(rowMin(distay[(nrow(data)+1):dim(distay)[1],1:nrow(data)]))
43  }
44  return(sy/(sx+sy))
45}

```

C.3.2. TD

```

1  TD = function(data,m){
2    S <- NULL
3    diss_data<-daisy(x = data, stand=FALSE, metric = "gower")
4    for (i in 1:m) {
5      S[i]<-0
6      clust <- pam(diss_data, diss=TRUE, k = i)
7      for (j in 1:i){
8        data.x<-data[c(which(clust$clustering==j)),]
9        gow_mat<-gower.dist(data.x, data.y=data[clust$medoids[j],])
10       S[i] = S[i] + sum(gow_mat)
11     }
12   }
13   return(plot(1:m, S, type='b', ylab='Total deviation', xlab='Number of clusters', main='TD
14   ', col='cornflowerblue' , pch=16))
}

```

C.3.3. Gower_pam_silhouette

```

1  Gower_pam_silhouette<-function(diss, m){
2    S<-NULL
3    for (i in 1:m){
4      clust<-pam(diss, diss = TRUE, i)
5      S[i]<-clust$silinfo$avg.width
6    }
7    return(plot(1:m, S, type='b', ylab='Average Silhouette Width', xlab='Number of clusters',
8    main='Silhouette', col='cornflowerblue' , pch=16))
}

```

C.3.4. diversity

```

1  diversity<-function(data, m){
2    Sr<-NULL
3    for (i in 1:m){
4      Sr[i]<-0
5      clust<-kproto(x=data, k=i, nstart = 5, verbose = FALSE)
6      TSumD<-sum(clust$dists)
7      WSumD<-clust$tot.withinss
8      BSumD<-TSumD-WSumD
9      Sr[i]<-BSumD/TSumD
10    }
11    return(plot(1:m, Sr, type='b', ylab='Quality(Diversity)', xlab='Number of clusters',
12    main='Quality', col='cornflowerblue' , pch=16))
13  }

```

C.3.5. TSD

```

1
2   TSD = function(data, m){
3     SD<- NULL
4     for (i in 1:m){
5       SD[i]<-0
6       kpres<-kproto(data, k = i, nstart = 5, verbose = FALSE)
7       for (j in 1:i){
8         SD[i]<-SD[i]+sum(kpres$dists[c(which(kpres$cluster==j)),j])
9       }
10     }
11     return(plot(1:m, SD, type='b', ylab='Total sum distances', xlab='Number of clusters',
12               main='TSD', col='cornflowerblue' , pch=16))
13   }

```

C.3.6. fuzzyalgo

```

1
2   segmax <-function(l){
3     #Función auxiliar a la función fuzzyalgo
4     # Calcula el segundo máximo
5     l<-l[-which.max(l)]
6     return(max(l))
7   }
8

```

```

1   fuzzyalgo <- function(membership){
2     Clust<-NA
3     for (i in 1:dim(membership)[1]) {
4       if(max(membership[i,])-segmax(membership[i,])<0.1){
5         Clust[i]<-dim(membership)[2]+1
6       }
7       else{
8         Clust[i]<-which.max(membership[i,])
9       }
10     }
11     return(Clust)
12   }

```

C.4. PLSDA

C.4.1. p.coef

```

1  p.coef<-function(pls_plsda,R, datosplsda, j){
2      #pls_plsda: modelo PLS o PLS-DA generado con la libreria rpls
3      #R: número de veces a repetir la prueba
4      #datosplsda: matriz de datos utilizada para crear el modelo
5      #j: posición en la matriz de datos de la variable respuesta
6      k=pls_plsda@summaryDF$pre
7      coefmod<-pls_plsda@coefficientMN
8      Y=datosplsda[,j]
9      a<-NULL
10     for (i in 1:R){
11         Yperm=sample(Y, replace=FALSE)
12         plsda.opls<-opls(datosplsda[,-j], factor(Yperm), scaleC='none', predI=k,
13             info.txtC='none', fig.pdfC='none', crossvalI=1)
14         a<-rbind(a,plsda.opls@coefficientMN)
15     }
16     a<-as.data.frame(a)
17     p.coefs<-matrix(2, nrow=length(plsda.opls@vipVn), ncol=2*length(plsda.opls@yMeanVn))
18     for(i in 1:length(plsda.opls@vipVn)){
19         coefs<-a[grep(rownames(a)[i],rownames(a)),]
20         for(l in 1:length(plsda.opls@yMeanVn)){
21             IC<-quantile(coefs[,l], prob=c(0.025,0.975))
22             pvalor<-mean(abs(coefs[,l])>abs(coefmod[i,l]))
23             p.coefs[i,2*l-1]<-pvalor
24             p.coefs[i,2*l]<-paste('(',IC[[1]],'-',IC[[2]],')')
25             rownames(p.coefs)<-names(plsda.opls@vipVn)
26         }
27     }
28     return(p.coefs)
29 }
```

C.4.2. plotweight

```
1
2 plotweight<-function(pls, name, axe1,axe2){
3     #hacer el gráfico a mano
4     # crear el plot
5     plot(pls@weightStarMN[,axe1], pls@weightStarMN[,axe2],
6         main = "Weights",
7         xlab = paste('w*c', axe1), ylab = paste('w*c', axe2),
8         pch = 18, col = "blue", xlim=c(-0.6,0.6) ,ylim=c(-0.6,0.6))
9
10    points(pls@cMN[,axe1], pls@cMN[,axe2], pch = 18, col = "red")
11    # Asignamos las etiquetas
12    text(pls@weightStarMN[,axe1], pls@weightStarMN[,axe2],
13        labels = row.names(pls@weightStarMN),
14        cex = 0.6, pos = 4, col = "black")
15    text(pls@cMN[,axe1], pls@cMN[,axe2], labels = c(paste(name, rownames(pls@cMN))), cex =
16        0.6, pos = 4, col = 'black')
17    abline(h=0, v=0)
18    legend("bottomright", legend = c("X", "Y"),
19        pch = 18, col = c("blue", "red"))
}
```

ANEXO D

Relación con los Objetivos de Desarrollo Sostenible

El presente Trabajo de Fin de Máster tiene un alto grado de relación con el tercer Objetivo de Desarrollo Sostenible (ODS) de la agenda 2030, el relacionado con la salud y el bienestar. En concreto con la meta de reducir la mortalidad prematura por enfermedades no transmisibles mediante la prevención y el tratamiento. Se relacionan con esta meta ya que, se ha conseguido dotar a los médicos que trabajan con pacientes de melanoma de una herramienta que les permita clasificar a los pacientes con etiopatogenia difusa y, por tanto, ayudarles en la elección de tratamientos más adecuados para los pacientes según su etiopatogenia.

Bibliografía

- Métodos jerárquicos de análisis cluster., 2022. URL <https://www.ugr.es/gallardo/pdf/cluster-3.pdf>.
- H. Abdi and L. J. Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.
- S. Baadel, F. Thabtah, and J. Lu. Overlapping clustering: A review. In *2016 SAI Computing Conference (SAI)*, pages 233–237, 2016. doi: 10.1109/SAI.2016.7555988.
- A. Banerjee and R. Dave. Validating clusters using the hopkins statistic. In *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, volume 1, pages 149–153 vol.1, 2004. doi: 10.1109/FUZZY.2004.1375706.
- M. Berwick, D. B. Buller, A. Cust, R. Gallagher, T. K. Lee, F. Meyskens, S. Pandey, N. E. Thomas, M. B. Veierød, and S. Ward. Melanoma epidemiology and prevention. *Cancer Treat Res*, 167:17–49, 2016.
- J. Camacho and A. Ferrer. Cross-validation in pca models with the element-wise k-fold (ekf) algorithm: Practical aspects. *Chemometrics and Intelligent Laboratory Systems*, 131:37–50, 2014. ISSN 0169-7439. doi: <https://doi.org/10.1016/j.chemolab.2013.12.003>. URL <https://www.sciencedirect.com/science/article/pii/S0169743913002335>.
- R. P. M. Casanova Seuma JM. Melanoma. *Atención Primaria*, 33(62.453):335–46, 2004.
- M. C. Cercato, E. Nagore, V. Ramazzotti, I. Sperduti, and C. Guillén. Improving sun-safe knowledge, attitude and behaviour in parents of primary school children: a pilot study. *J Cancer Educ*, 28(1):151–157, Mar. 2013.

- L. K. Dennis, M. J. Vanbeek, L. E. Beane Freeman, B. J. Smith, D. V. Dawson, and J. A. Coughlin. Sunburns and risk of cutaneous melanoma: does age matter? a comprehensive meta-analysis. *Ann Epidemiol*, 18(8):614–627, Aug. 2008.
- E. Diday and J. C. Simon. *Clustering Analysis*, pages 47–94. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976. ISBN 978-3-642-96303-2. doi: 10.1007/978-3-642-96303-2_3. URL https://doi.org/10.1007/978-3-642-96303-2_3.
- H. B. El-Serag and J. E. Everhart. Diabetes increases the risk of acute hepatic failure. *Gastroenterology*, 122(7):1822–1828, 2002.
- M. Ferraro, P. Giordani, and A. Serafini. fclust: An r package for fuzzy clustering. *The R Journal*, 11, 2019. URL <https://journal.r-project.org/archive/2019/RJ-2019-017/RJ-2019-017.pdf>.
- A. Galván. Consulta el libro blanco del cáncer de piel, Jun 2022. URL <https://aedv.es/consulta-libro-blanco-del-cancer-de-piel/>.
- J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.
- G. P. Guy, Jr, C. C. Thomas, T. Thompson, M. Watson, G. M. Massetti, L. C. Richardson, and Centers for Disease Control and Prevention (CDC). Vital signs: melanoma incidence and mortality trends and projections - united states, 1982-2030. *MMWR Morb Mortal Wkly Rep*, 64(21):591–596, June 2015.
- E. Hacker, E. Nagore, L. Cerroni, S. L. Woods, N. K. Hayward, B. Chapman, G. W. Montgomery, H. P. Soyer, and D. C. Whiteman. NRAS and BRAF mutations in cutaneous melanoma and the association with MC1R genotype: findings from spanish and austrian populations. *J Invest Dermatol*, 133(4):1027–1033, Oct. 2012.
- J. Han, M. Kamber, and J. Pei. *10 - Cluster Analysis: Basic Concepts and Methods*, pages 443 – 495. 12 2012. ISBN 9780123814791. doi: 10.1016/B978-0-12-381479-1.00010-1.
- J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics and Data Analysis*, 56(6):1869–1879, 2012. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2011.11.012>. URL <https://www.sciencedirect.com/science/article/pii/S0167947311004099>.

- A. Kassambara and F. Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020. URL <https://CRAN.R-project.org/package=factoextra>. R package version 1.0.7.
- A. Kassambara, M. Kosinski, and P. Biecek. *survminer: Drawing Survival Curves using 'ggplot2'*, 2021. URL <https://CRAN.R-project.org/package=survminer>. R package version 0.4.9.
- F. Kherif and A. Latypova. Principal component analysis. In *Machine Learning*, pages 209–225. Elsevier, 2020.
- D. G. Kleinbaum, M. Klein, et al. *Survival analysis: a self-learning text*, volume 3. Springer, 2012.
- G. Liszkay, Z. Kiss, R. Gyulai, J. Oláh, P. Holló, G. Emri, A. Csejtei, I. Kenessey, A. Benedek, Z. Polányi, Z. Nagy-Erdei, A. Daniel, K. Knollmajer, M. Várnai, Z. Vokó, B. Nagy, G. Rokszin, I. Fábián, Z. Barcza, and C. Polgár. Changing trends in melanoma incidence and decreasing melanoma mortality in hungary between 2011 and 2019: A nationwide epidemiological study. *Front Oncol*, 10:612459, Feb. 2021.
- A. M. Maceira, S. K. Prasad, P. N. Hawkins, M. Roughton, and D. J. Pennell. Cardiovascular magnetic resonance and prognosis in cardiac amyloidosis. *Journal of Cardiovascular Magnetic Resonance*, 10(1):1–11, 2008.
- T. S. Madhulatha. An overview on clustering methods. *CoRR*, abs/1205.1117, 2012. URL <http://arxiv.org/abs/1205.1117>.
- M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2021. URL <https://CRAN.R-project.org/package=cluster>. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source).
- G. W. Milligan and M. C. Cooper. Methodology review: Clustering methods. *Applied Psychological Measurement*, 11(4):329–354, 1987. doi: 10.1177/014662168701100401. URL <https://doi.org/10.1177/014662168701100401>.
- E. Nagore, V. Oliver, R. Botella-Estrada, S. Moreno-Picot, A. Insa, et al. Prognostic factors in localized invasive cutaneous melanoma: high value of mitotic rate, vascular invasion and microscopic satellitosis. *Melanoma research*, 15(3):169–177, 2005.

- E. Nagore, R. Botella-Estrada, C. Requena, C. Serra-Guillén, A. Martorell, L. Hueso, B. Llombart, O. Sanmartín, and C. Guillén. Clinical and epidemiologic profile of melanoma patients according to sun exposure of the tumor site. *Actas Dermo-Sifiliográficas*, 2009.
- G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.-D. Devignes, M. Kobayashi, O. Huttin, J. P. Ferreira, F. Zannad, P. Rossignol, and N. Girerd. Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Scientific Reports*, 11(1):4202, Feb. 2021.
- J. M. Ródenas, M. Delgado-Rodríguez, M. T. Herranz, J. Tercedor, and S. Serrano. Sun exposure, pigmentary traits, and risk of cutaneous malignant melanoma: a case-control study in a mediterranean population. *Cancer Causes Control*, 7(2):275–283, Mar. 1996.
- L. Rokach and O. Maimon. *Clustering Methods*, pages 321–352. Springer US, Boston, MA, 2005. ISBN 978-0-387-25465-4. doi: 10.1007/0-387-25465-X_15. URL https://doi.org/10.1007/0-387-25465-X_15.
- A. Rossi, M. Di Maio, P. Chiodini, R. M. Rudd, H. Okamoto, D. V. Skarlos, M. Fruh, W. Qian, T. Tamura, E. Samantas, et al. Carboplatin-or cisplatin-based chemotherapy in first-line treatment of small-cell lung cancer: the cocis meta-analysis of individual patient data. *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*, 2012.
- C. S. Punla, <https://orcid.org/0000-0002-1094-0018>, cspunla@bpsu.edu.ph, R. C. Farro, <https://orcid.org/0000-0002-3571-2716>, rcfarro@bpsu.edu.ph, and Bataan Peninsula State University Dinalupihan, Bataan, Philippines. Are we there yet?: An analysis of the competencies of BEED graduates of BPSU-DC. *International Multidisciplinary Research Journal*, 4(3):50–59, Sept. 2022.
- S. Sáenz, J. Conejo-Mir, and A. Cayuela. Melanoma epidemiology in spain. *Actas dermo-sifiliográficas*, 96(7):411—418, September 2005. ISSN 1578-2190. doi: 10.1016/s0001-7310(05)73105-7. URL [https://doi.org/10.1016/s0001-7310\(05\)73105-7](https://doi.org/10.1016/s0001-7310(05)73105-7).
- G. Szepannek. clustmixtype: User-friendly clustering of mixed-type data in r. *The R Journal*, pages 200–208, 2018. doi: 10.32614/RJ-2018-048. URL <https://doi.org/10.32614/RJ-2018-048>.

- Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.
- E. A. Thevenot, A. Roux, Y. Xu, E. Ezan, and C. Junot. Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and opls statistical analyses. *Journal of Proteome Research*, 14:3322–3335, 2015. URL <http://pubs.acs.org/doi/full/10.1021/acs.jproteome.5b00354>.
- C. R. UK. Melanoma skin cancer incidence statistics, 2022. URL <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer/incidence>.
- S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.
- S. van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6): 681–694, Mar. 1999. doi: 10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r. URL [https://doi.org/10.1002/\(sici\)1097-0258\(19990330\)18:6<681::aid-sim71>3.0.co;2-r](https://doi.org/10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r).
- J. N. Weinstein. A postgenomic visual icon. *Science*, 319(5871):1772–1773, 2008.
- E. W. Weisstein. Bonferroni correction. <https://mathworld.wolfram.com/>, 2004.
- Y. Xi, A. Formentini, M. Chien, D. B. Weir, J. J. Russo, J. Ju, M. Kornmann, and J. Ju. Prognostic values of micrornas in colorectal cancer. *Biomarker insights*, 1: 117727190600100009, 2006.