Universidad Zaragoza

# Trabajo Fin de Grado

Asociación de haplotipos mitocondriales con biomarcadores estructurales de MRI para la caracterización de la enfermedad de Alzheimer

Association of mitochondrial haplotypes with MRI structural biomarkers for the characterization of Alzheimer's disease


Autor

Juan Asensio Ayesa

Directoras

Elvira Mayordomo Cámara

Mónica Hernández Giménez



Grado en Ingeniería Informática -- Rama de Computación

Escuela de Ingeniería y Arquitectura - 2020

# Resumen

El desconocimiento de los factores que provocan la enfermedad de Alzheimer sigue siendo una dificultad para su diagnóstico. No obstante, existe un cierto consenso en que dicha enfermedad tiene un componente genético. Numerosos estudios tratan de encontrar relaciones entre variaciones en el genoma de los sujetos y marcadores indicativos del desarrollo de la enfermedad. Entre estos estudios existe un número de ellos que se centran en variaciones del ADN mitocondrial, aunque actualmente aún no existe un consenso generalizado sobre el papel que este tipo de ADN puede desempeñar en la enfermedad. En este Trabajo de Fin de Grado se ha realizado un estudio mediante diversas técnicas de aprendizaje automático tratando de replicar los resultados propuestos en el trabajo de investigación de P.G. Ridge. Dichas técnicas podrían ser divididas en modelos de selección de variables como Lasso, Elastic-net y Group Lasso que nos han permitido seleccionar aquellas variaciones genéticas que estén más relacionadas con la enfermedad y modelos de regresión. En este trabajo se ha utilizado el modelo de máquinas de vectores de soporte (SVM) como modelo de regresión. Mediante este modelo se ha podido estudiar la evolución del error en función de las variaciones incluidas en él. Estas dos clases de modelos, en su conjunto, han permitido evaluar las relaciones entre los datos genéticos utilizados y el fenotipo estudiado, que en este caso es la atrofia del hipocampo izquierdo. Los datos del fenotipo han sido extraídos de distintas lecturas de imagen por resonancia magnética (MRI) que miden el volumen del hipocampo, mientras que los datos genéticos pertenecen al genotipado del ADN mitocondrial de diversos sujetos. Estos datos han sido extraídos del portal ADNI (Alzheimer's Disease Neuroimaging Initiative), una iniciativa iniciada en 2004 con el objetivo de permitir a investigadores de todo el mundo compartir información con el fin de avanzar en la investigación de esta enfermedad.



## Agradecimientos

En primer lugar, agradecer a las tutoras Elvira y Mónica por su apoyo e interés a la hora ayudarme a la hora de elaborar este trabajo.

Este trabajo está dedicado a mis abuelos, mi familia biológica, la familia que conocí en Roma, mis amigos, cuyo apoyo ha sido indispensable en este 2020 y por último a todos los compañeros que me han acompañado durante estos 4 años compartiendo alegrías y frustraciones, siendo un orgullo haber trabajado a su lado.

# Introducción

En 1907 [1] Alois Alzheimer describió el caso de una mujer de 51 años la cual presentaba un deterioro de la memoria relativamente rápido junto con trastornos psiquiátricos. Murió cuatro años después. Pese a que en aquella época ya se conocían condiciones neurológicas crónicas que terminaban siendo fatales, la temprana edad de comienzo de esta condición así como el hallazgo de ovillos neurofibrilares en el cerebro (neurofibrillary tangle, NFT) hacían única a esta condición. La justificación del Alzheimer como una nueva enfermedad, así como las motivaciones del psiquiatra Emil Kraepelin de promoverla como una nueva condición continúan debatiéndose.

Actualmente, la probable enfermedad de Alzheimer se diagnostica aproximadamente en el 70% de los casos de demencia. La incidencia de esta enfermedad aumenta con la edad, doblándose cada 5 a 10 años. La prevalencia también aumenta con la edad, en este caso de forma exponencial. Además de la edad existen otros factores de riesgo como el sexo, nivel de educación, enfermedades vasculares, factores ambientales y genéticos o el historial familiar, siendo especialmente frecuente la transmisión materna [2].

En el futuro se prevé que, debido a la baja natalidad y al aumento de la esperanza de vida se produzca un envejecimiento de la población. Teniendo en cuenta el hecho de que la edad es el factor de riesgo más significativo para la enfermedad de Alzheimer, no se pueden ignorar los costes humanos y económicos que esta enfermedad puede representar en las siguientes décadas.

Pese a esto y a los continuos esfuerzos científicos aún no se ha encontrado una cura para la enfermedad, aunque sí que se han realizado estudios que parecen demostrar que un estilo de vida saludable [3] o la actividad intelectual [4] de las personas que sufren Alzheimer puede retrasar la aparición de la demencia, la cual consiste en un deterioro grave de la capacidad mental que interfiere con la vida cotidiana. Aunque no exista una cura, sí se han desarrollado una serie de tratamientos que buscan mejorar el nivel de vida de las personas que padecen la enfermedad así como retrasar la aparición de la demencia. Para que esos tratamientos tengan el efecto deseado es importante una detección temprana de la enfermedad. Existe un importante esfuerzo científico centrado, precisamente, en la detección temprana de la enfermedad, lo que ha dado lugar a la aparición de iniciativas como la Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI se inició en 2004, bajo el liderazgo del Dr. Michael W. Weiner, contando con subvenciones tanto de empresas privadas como de organismos públicos. ADNI comenzó con un estudio de 5 años de duración denominado ADNI1, el cual fue ampliado durante 2 años en 2009 (ADNI-GO) y en 2011 y 2016 (ADNI-2 y ADNI-3). En cada fase del estudio se han mejorado los protocolos de adquisición de datos y reclutado nuevos pacientes. Desde su lanzamiento ADNI ha hecho importantes contribuciones a la investigación del Alzheimer, permitiendo el intercambio de información de investigadores alrededor del mundo y favoreciendo la comprensión de muy diversos aspectos de la enfermedad.

La fuerte heredabilidad de la enfermedad de Alzheimer, especialmente la transmisión materna, así como la aparición de déficits de Citocromo c oxidasa, fundamental en el ciclo de producción de energía en organismos aeróbicos, en personas con Alzheimer, han propiciado la aparición de un creciente número de estudios tratando de explicar relaciones entre polimorfismos que afectan a un único nucleótido, denominados en inglés single nucleotide polymorphism (SNP), y determinados marcadores de la enfermedad de Alzheimer. Estas variaciones en el genotipo se obtienen mediante un proceso denominado genotipado. Se entiende como genotipo la información genética que tiene un organismo. Dicha información viene codificada en largas cadenas de proteínas denominadas ADN [23]. La estructura de estas cadenas está compuesta por una doble hélice, cada una de esas hebras está formada por unos compuestos denominados nucleótidos. Los nucleótidos presentan una parte formada por un monosacárido y un grupo fosfato, lo que constituiría la "columna vertebral" de la doble hélice, y por una base nitrogenada, que es lo que permite la interacción entre las hebras del ADN. Dichas bases son: Adenina (A), Citosina (C), Guanina (G) y Timina (T), de forma que únicamente se pueden realizar enlaces entre

A-T y G-C, en la figura a continuación se muestra la estructura del ADN. El orden de combinación de estos nucleótidos es lo que codifica la información. El ARN, en cambio, está conformado por una única hebra y es utilizado para transportar la información del ADN a través de la célula para completar procesos biológicos.

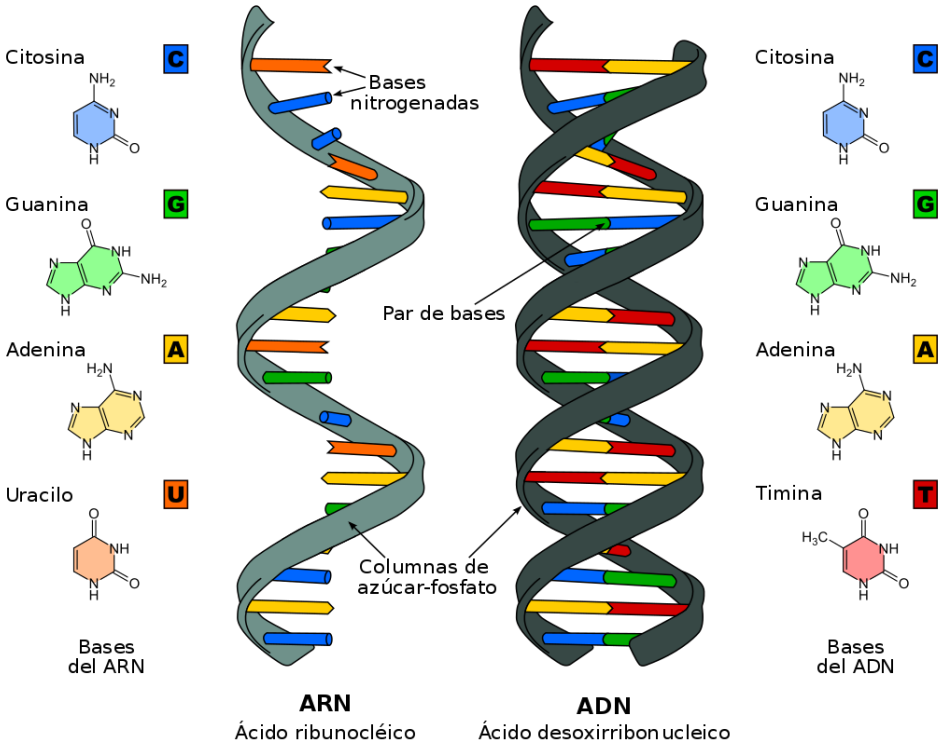


Figura 1.1 Estructura del ADN y del ARN

En el caso de los humanos existe tanto el ADN celular o nuclear como el ADN mitocondrial. Existen varias diferencias entre el ADN nuclear y el ADN mitocondrial. La más notable es que el ADN nuclear está compuesto por alrededor de 3200 millones de bases, agrupados en alrededor de 25000 genes, mientras que el ADN mitocondrial está compuesto por 16569 pares y 37 genes. Los genes son secciones del ADN que proporcionan a la célula la información para crear un determinado producto génico, que puede ser bien una proteína o ARN. En el ADN nuclear se encuentran agrupados en cromosomas, en los cuales cada gen ocupa un locus. Un individuo puede tener distintas versiones de un mismo gen denominado alelo. En el caso del ADN mitocondrial, éste, al igual que el ADN bacteriano está dispuesto en una forma circular, y no se presentan distintas versiones de un único gen ya que se hereda únicamente de la madre biológica.

El proceso de genotipado consiste entonces en la obtención del genotipo de un organismo biológico en concreto. Esto permite identificar posibles mutaciones presentes en un individuo. Las mutaciones son variaciones de la secuencia genética, las cuales pueden ser producidas de forma natural durante el proceso de división celular. Como se ha comentado anteriormente, el ADN ocupa un papel fundamental en diversos procesos biológicos, por lo que una variación en éste puede llevar al mal funcionamiento de uno de estos procesos, y, eventualmente a la aparición de una enfermedad.

Una posible forma de evaluar si una variación genética está relacionada con la aparición de una enfermedad es analizar si está relacionada con algún cambio en un fenotipo del sujeto que pueda indicar la aparición de dicha enfermedad. El fenotipo consiste en el conjunto de rasgos tanto físicos como conductuales fruto de la expresión del genotipo en función de un determinado entorno.

Existen varios estudios que sitúan al hipocampo como una de las primeras regiones del cerebro que sufren daño debido a la enfermedad de Alzheimer [29] [30], por lo que, junto con el amplio consenso en las funciones del hipocampo, lo sitúan como blanco de estudio fenotípico para comprender la enfermedad. El

hipocampo [27] es una región del cerebro que fue descrita por primera vez en el siglo XVI por el anatomista Giulio Cesare Aranzio. El hipocampo se encuentra en la parte medial del lóbulo temporal del cerebro.

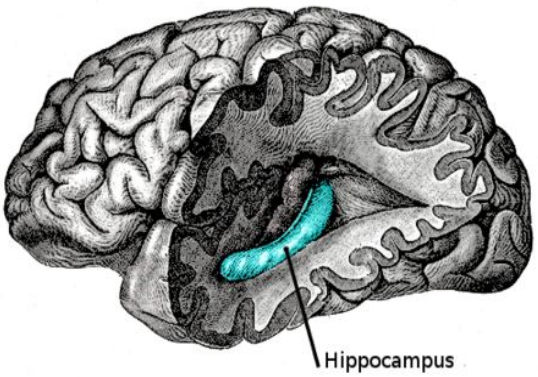


Figura 1.2 Situación del hipocampo en el cerebro

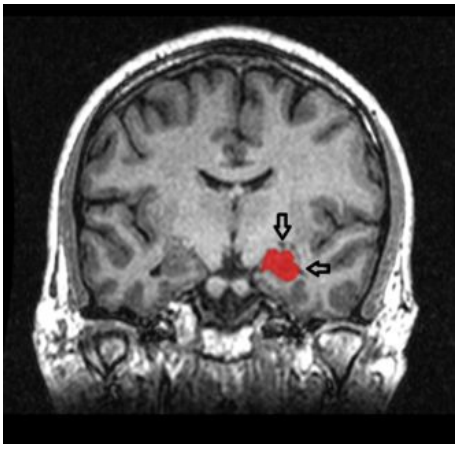

Figura 1.3 Hipocampo en una resonancia magnética

Históricamente se asociaba al hipocampo con el olfato. Esta teoría ha acabado siendo descartada y actualmente existen 3 distintas líneas teóricas que relacionan al hipocampo con la inhibición, la memoria y la orientación. La primera línea teórica justificaba la relación del hipocampo con la inhibición con dos observaciones en animales: la primera, que los animales con lesiones en el hipocampo tienden a ser hiperactivos y la segunda, que los animales sometidos a estas lesiones también tienen problemas para inhibir respuestas que previamente se les había enseñado. La segunda línea, aunque con precedentes históricos, presentó un fuerte impulso tras el artículo de Scoville y Brenda Milner [28] que explicaba cómo tras la destrucción quirúrgica del hipocampo con el objetivo de evitar los ataques de epilepsia el sujeto presentaba una grave amnesia. Finalmente, la función en la orientación del hipocampo está basada en las teorías de "Mapas cognitivos" de E.C. Tolman's las cuales fueron una influencia para O'Keefe y Nadel. O'Keefe junto con su alumno Dostrovsky descubrieron en 1971 neuronas en el hipocampo de ratas que parecían tener relación con la localización de la rata en el espacio. Aunque en un inicio este estudio causó escepticismo, en la actualidad existe un consenso prácticamente universal acerca del papel del hipocampo en las funciones de memoria y localización espacial, aunque los detalles están aún sometidos a debate.

Las técnicas de neuroimagen actuales permiten obtener una medida más directa del impacto de las variaciones genéticas a nivel anatómico. Esto ha dado lugar al uso de métodos como la obtención de imágenes por resonancia magnética (MRI por sus siglas en inglés) que han permitido caracterizar ritmos

acelerados de atrofia en el volumen de las diferentes regiones cerebrales, establecido como un indicador de riesgo para el desarrollo de la enfermedad de Alzheimer [5].
No obstante, estos métodos de diagnóstico se ven obstaculizados por el hecho de que, aunque existan numerosas hipótesis sobre ello, la causa exacta de la enfermedad de Alzheimer, tal y como se ha mencionado anteriormente es desconocida [45]. La hipótesis más aceptada es la hipótesis de la cascada amiloidea [34], por la que la proteína precursora amiloidea (APP por sus siglas en inglés) es dividida por distintas secretasas (α,β y γ) formando péptidos, uno de ellos es $A\beta_{42}$. Estos compuestos pueden acabar formando fibrillas extracelulares dispuestas en láminas β (una de las posibles disposiciones adoptadas por las proteínas). Estas láminas forman parte de las placas seniles [35]. Las placas seniles son depósitos extracelulares de compuestos beta-amiloide en la sustancia gris del cerebro [36], y se cree que es el primer paso en el desarrollo de la enfermedad del Alzheimer [37].
En cuanto a factores genéticos, se sabe que varios genes o haplotipos albergan variantes causales o de riesgo para la enfermedad de Alzheimer. Estos factores son distintos para los dos tipos principales de Alzheimer definidos por la edad de aparición. El primer tipo es el Alzheimer de inicio temprano (EOAD por sus siglas en inglés) y el segundo el Alzheimer de inicio tardío (LOAD por sus siglas en inglés). Se conoce que las mutaciones en tres genes causan la enfermedad de Alzheimer de inicio temprano, dichos genes son: proteína precursora amiloidea APP [38], presenilina-1 (PSEN1) [39], presenilina-2(PSEN2) [40]. Por otra parte para LOAD, pese a que se han identificado numerosos factores de riesgo genéticos, no se ha identificado ningún gen causante de este tipo de Alzheimer. No obstante se ha demostrado que, por ejemplo, el gen APOE presenta una relación con la enfermedad [41], aunque su importancia y rol se desconocen por ahora. Dicho gen está relacionado con la síntesis de apolipoproteinas, al igual que el gen CLU, una variante de este gen (rs11136000) se ha identificado en varias ocasiones como un alelo protector [42-44].
Tal y como se ha explicado anteriormente, es bien conocido el mal funcionamiento de las mitocondrias en la enfermedad de Alzheimer. Si este mal funcionamiento es una causa o un efecto del Alzheimer se desconoce, de la misma forma que el rol, si es que existe, del genoma mitocondrial como factor de riesgo de la enfermedad del Alzheimer es desconocido. Numerosos estudios (Figura 4.5.1) han tratado de encontrar relaciones entre variaciones del genoma mitocondrial, o haplotipos con la enfermedad de Alzheimer. Pese a que algunos de estos estudios han identificado asociaciones significativas no existe un consenso al respecto, incluso existen estudios que presentan resultados contradictorios.
En ese sentido son interesantes las conclusiones del reciente estado del arte [45]:

*Numerous mitochondrial haplogroups and SNPs have been reported to influence risk for AD, but the majority of these have not been replicated, nor experimentally validated. The role of the mitochondrial genome in AD remains elusive, and several impediments exist to fully understand the relationship between the mitochondrial genome and AD. Yet, by leveraging existing datasets and implementing appropriate analysis approaches, determining the role of mitochondrial genetics in risk for AD is possible.*

Este Trabajo de Final de Grado (TFG) tiene entonces como objetivo realizar un análisis de las posibles relaciones existentes entre el fenotipo de los sujetos y su genoma mitocondrial mediante el uso de diversas técnicas de aprendizaje automático, las cuales se pueden separar en modelos de selección de variables, como Lasso, Elastic net o Group Lasso o modelos de regresión como SVM, tratando de replicar los resultados propuestos por P.G Ridge en el artículo Mitochondrial Genetics of Alzheimer's Disease and Aging publicado en 2013 [5]. En este TFG se han utilizado implementaciones realizadas por el alumno Eduardo Alonso en su TFG de 2020 [31], concretamente herramientas de preprocesamiento de datos genéticos nucleares así como sus resultados obtenidos de su implementación de los modelos Lasso y Elastic net, con el fin de poder comparar los resultados obtenidos con ADN mitocondrial y ADN nuclear,

así como de poder validar ciertos métodos al poder comparar los resultados con los obtenidos mediante otra implementación.

En la Sección 2 se detallan los métodos utilizados en este TFG. En la Sección 3 se describen los materiales clínicos utilizados en este trabajo. En la Sección 4 se presentan los resultados más relevantes de todos los obtenidos. Finalmente, la Sección 5 presenta las conclusiones de este TFG y sus posibles extensiones en líneas futuras de investigación.

# Métodos

## 2.1 Lasso

Lasso fue propuesto en 1996 por Tibshirani [6] como un método de estimación de modelos lineales. Lasso minimiza la suma de cuadrados residual del valor absoluto de los coeficientes de forma que ésta sea menor que una constante. Esta restricción provoca que los coeficientes tiendan a tomar el valor de 0, lo que permite obtener modelos interpretables. En aquella época Ridge regression era uno de los métodos más utilizados para reducir el valor de los coeficientes reduciendo así el overfitting. Como Ridge regression no realiza ningún tipo de selección de variables, se resolvía este problema utilizando un método denominado stepwise regression [18]. Stepwise regression consiste en realizar iteraciones seleccionando si una variable pertenece o no al modelo. Esto solo mejora la predicción en los casos en los que existan pocas variables que tengan una relación fuerte con la variable independiente, en el resto de los casos puede empeorar el error de predicción. Tibshirani demostró en su artículo mediante simulaciones que Lasso permitía realizar selección de variables para generar modelos interpretables a la vez que mantenía la estabilidad de Ridge regression.
Considerando la típica situación de regresión donde tenemos $(X_i, y_i)\ i = 1, 2, 3, ..., N$ observaciones donde $X_i = (x_{i1}, ... , x_{ip})^T$ y $y_i$ son las características y valor para la observación i, Lasso tiene como objetivo resolver :

$$min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \ sujeto\ a \ \sum_{j=1}^{N} \left|\beta_j\right| \leq t.$$

Siendo el valor de t la constante mencionada anteriormente. La expresión además se puede escribir de forma más compacta:

$$min_{\beta_0,\beta} \left\{ \frac{1}{N} \left|\left| y - \beta_0 - X\beta \right|\right|_2^2 \right\} \ sujeto\ a \ ||\beta||_1 \leq t.$$

Si denotamos la media escalar de los datos $x_i$ como $\overline{x}$ y de forma similar con $y_i$ tenemos que la estimación resultante para $\beta_0$ termina siendo $\widehat{\beta_0} = \overline{y} - \overline{x_i}^T \beta$ por lo que es estándar trabajar con variables centralizadas, adicionalmente las variables son estandarizadas.
El problema entonces se puede reescribir en su forma Lagrangiana:

$$min_{\beta \in \Re^p} \left\{ \frac{1}{N} ||y - X\beta||_2^2 + \alpha ||\beta||_1 \right\}$$

donde la relación entre $\alpha$ y t dependerá de los datos, además $\alpha$ determinará el grado de "sparsity" de la solución, es decir, el número de características seleccionadas por el modelo.

### 2.1.1 Implementación

Se ha utilizado la implementación de la librería de Python (Scikit-learn) de este modelo [7]. Dicha implementación minimiza la función:

$$\frac{1}{2N} \cdot ||y - Xw||_2^2 + \alpha||w||_1$$

que corresponde a la forma Lagrangiana del problema de Lasso. La variable w corresponde a los pesos de las distintas características. En cuanto a la centralización de los datos, el modelo implementado en Python por defecto no supone que los datos estén centrados.

En el contexto de este TFG, $X$ es la matriz con los SNP que presenta cada sujeto mientras que $y$ es un fenotipo relacionado con el diagnóstico de probable enfermedad de Alzheimer. El modelo entonces adjudicará ciertos pesos a cada uno de los SNPs , lo que se podrá utilizar como indicador de qué SNPs son más relevantes para el fenotipo en cuestión. Explicaremos la metodología utilizada con más detalle en la sección 2.6.

## 2.2 Elastic net

En el año 2005, Hui Zou y Trevor Hastie proponían un nuevo método nuevo de regularización y selección de variables denominado elastic net [8]. En el artículo presentan pruebas realizadas con datos reales y simulación que demuestran que elastic net suele presentar mejores resultados que Lasso, a la vez que sigue dotando de una solución *sparse* , es decir, que realiza selección de variables para crear un modelo interpretable. Además elastic net favorece un efecto de agrupamiento de tal forma que variables fuertemente relacionadas se quedan dentro o fuera del modelo en conjunto. Elastic net es también particularmente efectivo cuando el número de variables o características (p) es mucho mayor que el de observaciones (n), mientras que Lasso no obtiene resultados muy buenos en el caso de $p >> n$. Lasso también tiene problemas con variables fuertemente relacionadas ya que el modelo solo seleccionaría una variable de ese grupo en lugar de todas. Elastic net soluciona estos problemas de Lasso a la vez que mantiene sus ventajas.

Volviendo a la típica situación de regresión donde tenemos $(X_i, y_i)\ i = 1, 2, 3, ..., N$ observaciones donde $X_i = (x_{i1}, ... , x_{ip})^T$ e $y_i$ son las características y valor para la observación i (se supone que los datos están centrados y estandarizados), se define como *naive elastic net* a:

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1$$

Siendo $\lambda_1$ y $\lambda_2$ cualquier valor no negativo. Elastic net trata de resolver:

$$\widehat{\beta} = argmin_\beta \{L(\lambda_1, \lambda_2, \beta)\}$$

Elastic net es entonces, un método de regresión que utiliza tanto la regularización $L_1$ como $L_2$, lo que permite generar un modelo con pocas variables seleccionadas al igual que Lasso mientras que se mantienen las propiedades de regularización de Ridge.

### 2.2.1 Implementación

Se ha utilizado la implementación de la librería de Python (Scikit-learn) de este modelo [9] de elastic net. La versión implementada en Python aprovecha un trading-off entre Lasso y Ridge, tal y como se ha comentado en el apartado anterior.

La función a minimizar por la librería de Python es:

$$min_w \frac{1}{2N}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

donde ρ representa la variable L1-ratio de la función de Python, que, como se puede observar es el encargado de controlar el ratio entre la regularización $L_1$ y $L_2$.

En el contexto de este TFG, elastic net permitirá seleccionar los SNPs relevantes a la variable dependiente y, de forma similar a Lasso. Explicaremos la metodología utilizada con más detalle en la sección 2.6.

## 2.3 Sparse Group Lasso

En algunos problemas de regresión con alta dimensionalidad puede ser útil realizar determinadas suposiciones sobre los datos de entrenamiento del modelo. Como se ha mencionado en los apartados anteriores este era uno de los puntos débiles de Lasso porque únicamente seleccionaba una de las variables pertenecientes a ese grupo mientras que Elastic net obtenía mejores resultados. En 2013 Noah Simon, Jerome Friedman, Trevor Hastie y Rob Tibshirani propusieron el método de Sparse Group Lasso [10] para problemas en los que las variables estén agrupadas y presenten un comportamiento sparse dentro y fuera de cada grupo.

Volviendo al caso de un problema de regresión de los puntos 2.1 y 2.2 y tomando la función de coste del modelo Lasso:

$$min_{\beta \in \Re^p} \left\{ \frac{1}{N}||y - X\beta||_2^2 + \alpha||\beta||_1 \right\}$$

Se puede dar el caso en el que nuestras variables de X estén divididas en varios grupos. En estos casos nos puede interesar no sólo que se seleccionen pocas variables de X en general sino además que se seleccionen entre un número limitado de grupos., para esto Yuan y Lin propusieron en 2006 Group Lasso:

$$min_\beta \frac{1}{2}\left|\left|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right|\right|_2^2 + \lambda \sum_{i=1}^{m} \sqrt{p_l}\left|\left|\beta^{(l)}\right|\right|_2$$

Siendo $X^{(l)}$ la submatriz de $X$ cuyas columnas corresponden a los elementos del grupo $l$, $\beta^{(l)}$ el vector de coeficientes de ese grupo y $p_l$ la longitud del grupo $l$. Este modelo proporciona un conjunto reducido de grupos, no obstante si un grupo es seleccionado todos los elementos de ese grupo tendrán un coeficiente distinto de cero, pero en algunos casos se puede necesitar elegir elementos particularmente importantes de ciertos grupos, como por ejemplo un SNP de un gen en determinado. Por esto es por lo que los autores proponen el siguiente modelo:

$$min_\beta \frac{1}{2N}\left|\left|y - \sum_{l=1}^{m} X^{(l)}\beta^{(l)}\right|\right|_2^2 + (1-\alpha)\lambda \sum_{i=1}^{m} \sqrt{p_l}\left|\left|\beta^{(l)}\right|\right|_2 + \alpha\lambda||\beta||_1$$

El modelo representa una combinación entre la solución de Lasso y Group Lasso de forma que $\alpha = 0$ proporciona una solución de tipo Group Lasso y $\alpha = 1$ una de tipo Lasso.

### 2.3.1 Implementación

Existe una API implementada en Python del modelo Sparse Group Lasso [12].

## 2.4 SVM

SVM se corresponde con las siglas en inglés de *Support Vector Machines*, una técnica desarrollada en AT&T Bell Laboratories, que es considerado uno de los modelos de predicción más robustos basado en aprendizaje estadístico. SVM es comúnmente utilizado en problemas de clasificación, de forma que cada uno de los vectores de soporte genera un hiperplano que permite separar las distintas clases. En algunos problemas no es posible realizar una separación lineal entre las distintas clases, por lo que se utilizan los denominados *kernels*, los cuales permiten mapear los datos de entrada en espacios de alta dimensionalidad, de forma que en ese superespacio las clases puedan ser linealmente separables. La figura 2.4.1 muestra un ejemplo de este fenómeno. Los puntos $\boldsymbol{x} = \{x_1, x_2\}$ no son linealmente separables pero se puede aplicar una función de transformación (kernel) tal que: $\Phi(x) = x_1^2, x_2^2, \sqrt{2}x_1x_2$ de forma que los datos pasan a estar proyectados a un espacio tridimensional y podrían ser separados por un hiperplano de 2 dimensiones [16]. La idea de regresión lineal con SVM (SVR) es similar a la utilizada en problemas de clasificación, mapear los datos de entrada a un espacio de altas dimensiones de forma que se pueda realizar una regresión lineal.

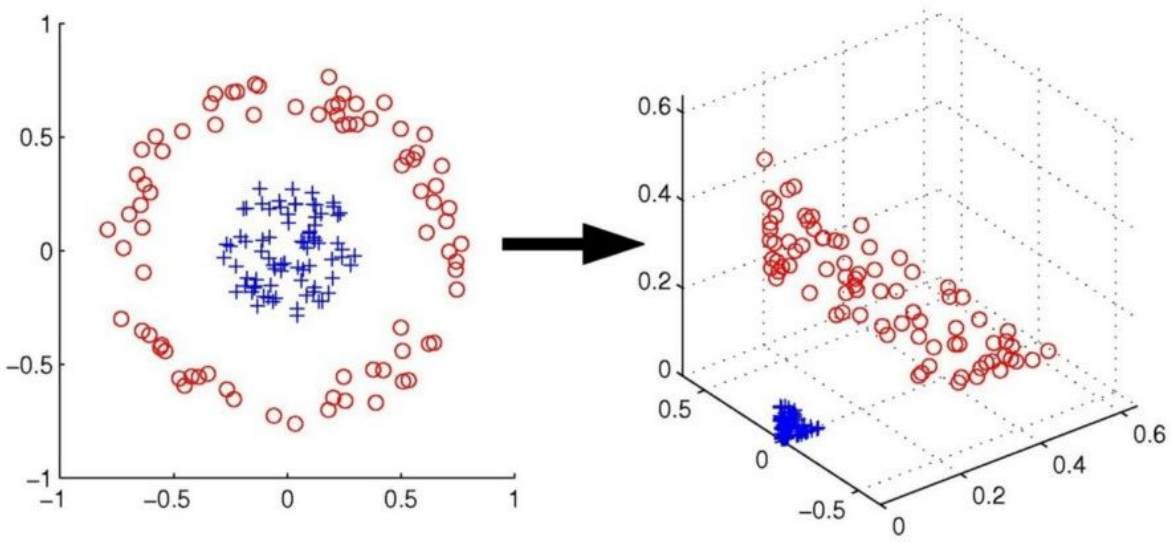


Figura 2.4.1 Ejemplo de uso de un kernel

### 2.4.1 Implementación

La librería scikit de Python proporciona una API para SVM [17]. Dicha implementación resuelve el siguiente problema:

$$min_{w,b,\zeta,\zeta^*} \ \frac{1}{2}w^T w \ + \ C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*)$$

$$sujeto\ a \quad y_i \ - \ w^T \Phi(x_i) \ - \ b \ \leq \ \varepsilon \ + \ \zeta_i,$$

$$w^T \phi(x_i) \ + \ b \ - y_i \leq \varepsilon + \zeta_i^*,$$

$$\zeta_i, \zeta_i^* \geq 0, i \ = \ 1, ..., n$$

De esta forma se penalizan las muestras cuya predicción tiene al menos ε de error de su objetivo.

En este caso se han utilizado los kernel polinomial y lineal de la librería mencionada.

## 2.5 Selección de SNP relevantes

### 2.5.1 Trabajo de P.G. Ridge

En el artículo de P.G. Ridge et al [5] se propone el uso del programa propuesto por Templeton, Crandall, and Sing (conocido como TCS por las iniciales de los apellidos de sus autores) [19] Para estimar una red de haplotipos utilizando los datos genéticos que se especificarán en la sección 3. Esta red de haplotipos constituía un marco de referencia del cual seleccionar haplotipos relacionados y agruparlos para realizar una asociación entre fenotipo y genotipo. Dicha asociación se lleva a cabo mediante una técnica denominada TreeScanning [20]. Dicho análisis se llevó a cabo con la herramienta TreeScan [21]. Esa herramienta se encontraba disponible en la página de darwin.uvigo.es , al igual que la herramienta TCS, pero en el momento de realización de este trabajo ya no se encontraba en dicha página. Se contactó con David Posada (UVIGO), miembro del equipo de creación de dicha herramienta. Esta persona nos informó de que no conocía el estado de esa herramienta ya que llevaba un tiempo sin mantenerla y que tampoco contaba con ninguna versión disponible. Durante la realización de este TFG se estudió la opción de implementar una versión de TreeScanning siguiendo [20]. La descripción propuesta en el artículo es la siguiente:

*"In the first round of tree scanning, a branch is cut in the haplotype tree. All of the haplotypes on one side of the cut are grouped together and treated as a single allele, say A. All the haplotypes on the other side of the cut are grouped together and treated as a single allele, say B. These two alleles define three potential genotypes: AA, AB, and BB. Associations between phenotypes and these genotypes are measured by the F-statistic from a standard one-way ANOVA."*

De este estudio se llegó a la conclusión de que al realizar una implementación propia de la herramienta se podrían producir variaciones de relevancia en la salida del algoritmo, como por ejemplo la forma de resolver todos los empates a la hora de realizar los cortes, que impidiesen comparar los resultados con los obtenidos por P.G. Ridge, lo que degradaría enormemente la capacidad de reproducir exactamente el estudio original. Por ello, se decidió aproximar el problema utilizando técnicas de aprendizaje automático propuestas con éxito en la literatura para ADN nuclear y estudiar la reproducibilidad de los resultados de P.G. Ridge con estas metodologías de referencia para este tipo de información genética.

### 2.5.2 Lasso / Elastic Net

Para elegir los SNP más relevantes con los métodos de Lasso y Elastic Net, se realizan N repeticiones dividiendo en cada una de ellas los datos de test y entrenamiento de forma aleatoria. En cada iteración se realiza un k-fold para elegir los mejores parámetros del modelo y se entrena con los datos de test para obtener los coeficientes del modelo para cada característica. Después, se calcula la media para cada coeficiente. El algoritmo se detalla en la figura 2.6.2.1. Tras ejecutar el algoritmo se ordenan las características para obtener el valor medio de sus coeficientes.

```
coefs = zeros(N_features)
for i in [1...N]:
        data_train,data_test = split(data,i)
        best_params = K_fold(model,data_train)
        best_model = train(model,data_test,best_params)
        coefs_i = best_model.coefs
        coefs = coefs + coefs_i
coefs = coefs / N
```

Figura 2.6.2.1 Descripción del algoritmo

#### 2.5.2.1 Hiperparámetros para Lasso

Como se muestra en el apartado 2.1.1 el parámetro α regula la penalización a los pesos del modelo, por lo que, intuitivamente, cuanto mayor sea el valor de este parámetro menos características serán seleccionadas para el modelo. Se probaron dos métodos para elegir el valor óptimo de α. El primero consistió en elegir los valores entre un rango amplio de magnitudes y el segundo en utilizar los valores de α utilizados por el alumno Eduardo Alonso [31]. Pese a que el orden de magnitud del valor óptimo de α pasaba de ser 1 en el primer caso a $10^{-3}$ en el segundo, los SNPs seleccionados con ambas opciones resultaron ser muy similares. Por ello, se optó por elegir el método que proporcionaba menores valores de α ya que esto permitía al modelo asignar valores no nulos a un mayor número de SNPs.

#### 2.5.2.2 Hiperparámetros para Elastic net

En elastic net hay que elegir el valor del parámetro α y de l1_ratio. Los posibles valores de α son los mismos que en el caso Lasso, mientras que l1_ratio, tal y como se indica en el punto 2.2 tiene que tomar un valor en 0 y 1, el valor es elegido entre 0.1 y 0.8, ya que un valor de 0 supondría una regresión Ridge y un valor de 1 sería el caso de Lasso. En este caso se quiere estudiar una combinación entre las dos.

### 2.5.3 Group Lasso

Para Group Lasso, se seleccionan los grupos que más veces han sido seleccionados, siguiendo el modelo comentado en la figura 2.6.2.1. En cuanto al agrupamiento de los SNP, el ADN mitocondrial cuenta con 37 genes cuya posición es conocida, por lo que basta con asignar a cada SNP un gen según la posición en la que se presenta.

#### 2.5.3.1 Hiperparámetros para Group Lasso

Una vez más se utilizan los mismos valores de α, mientras que los valores de la penalización para los grupos se eligen entre valores de orden $10^{-5}$ hasta  valores del orden de $10^{1}$ para comprobar si es efectivo agrupar los datos de entrada. Como se explicará en secciones posteriores se obtienen resultados que indican que agrupar los datos no tiene un efecto positivo en este caso.

## 2.5.4 SVM

En el apartado anterior se ha explicado cómo se han elegido los SNPs utilizando los modelos de Lasso y Elastic net para compararlos con los resultados propuestos por P.G. Ridge. Estos métodos son de manera natural métodos de selección de características. Con SVM se resuelve un problema de regresión, por lo que se ha tenido que enfocar el problema de manera distinta. El artículo de P.G. Ridge [5] proporciona unos resultados en los que ordena los SNPs según su p-valor obtenido, por lo que se ha decidido realizar una clase de *"stepwise regression"* simplificada, de forma que se iban añadiendo los SNP al modelo para comprobar la evolución del error según el p-valor de los SNP añadidos.
Los SNP se incluyen al modelo siguiendo tres órdenes distintos: según el orden de la lista de importancia de P.G. Ridge, en orden inverso, y en orden aleatorio. De esta forma, se ha buscado comprobar si efectivamente existe relación entre los p-valores y el rendimiento del modelo, en lugar de que las posibles variaciones del error se debieran a otras razones como por ejemplo, el número de características del modelo. Una descripción de este algoritmo se encuentra en la siguiente figura.

```
SNP = [snp_1,...,snp_m]
SNP_inverse = reverse(SNP)
SNP_random = rand_perm(SNP)
data_train,data_test = split(data)
for i in [1,...,M]:
        model_sorted = train(data_train,SNP[1:i])
        model_inverse = train(data_train,SNP_inverse[1:i])
        model_random = train(data_train,SNP_random[1:i])
        //Cada modelo se entrena únicamente con los SNP que le  corresponde
        prediction_sorted = predict(model_sorted,data_test)
        prediction_inverse= predict(model_inverse,data_test)
        prediction_random = predict(model_random,data_test)
        //Calcular errores de predicción de cada uno de los modelos
```

Figura 2.6.3.1 Algoritmo "stepwise" realizado con los SNP

Cabe destacar que antes de realizar el algoritmo previamente descrito hay que realizar un k-fold para seleccionar los parámetros adecuados para el entrenamiento. Hay que asegurarse que el algoritmo anterior utilice como conjunto de entrenamiento los mismos datos que se utilizan en el k-fold, es decir que el conjunto de test no aparezca en el k-fold. Esto se consigue utilizando la misma semilla (*seed)* para la función split de scikit de Python.

### 2.5.4.1 Hiperparámetros para SVM

En los modelos SVM hay que elegir los siguientes parámetros:

- C: La fuerza de la regularización es inversamente proporcional a su valor.
  - Se elige con valores desde $10^{-5}$ hasta 100
- degree: El grado del polinomio ( en caso de kernel polinomial)
  - Posibles valores: 2,3 o 4
- coef0: Término independiente del modelo
  - Tomado entre 0 y la media de los valores a predecir
- epsilon: Error admitido por el modelo

- Tomado entre 0 y la media de los valores a predecir

### 2.6.5 Estudio de frecuencias

Eduardo Alonso en su TFG [31] realiza un estudio de selección de frecuencias que consiste en el siguiente algoritmo:

```
apariciones = zeros(N_features)
for i in [1...N]:
        data_train,data_test = split(data,i)
        best_params = K_fold(model,data_train)
        best_model = train(model,data_test,best_params)
        coefs_i = best_model.coefs
        apariciones_i= coefs_i != 0
        apariciones = apariciones + apariciones_i
frecuencias = apariciones / N
```

Figura 2.6.5.1 Algoritmo para el estudio de la frecuencia de selección de SNPs

Se puede apreciar que el algoritmo anterior es muy similar al mostrado en el punto 2.6.2, pero en lugar de evaluar la media de los coeficientes, contabiliza las veces que el coeficiente para un determinado SNP es distinto de 0. Además existía una diferencia entre las dos implementaciones a la hora de seleccionar el mejor valor de α para entrenar con los datos de test. Eduardo Alonso [31] seleccionaba los parámetros que aseguraban que se seleccionan un mayor número de SNP, mientras que en este trabajo se decidió elegir el valor de los parámetros que generaba un menor error. Este método se implementó para comparar los resultados con los obtenidos por Eduardo Alonso [31], cambiando la política de selección de mejores parámetros para que se minimice el error.

### 2.6.6 Selección de hiperparámetros

En todos los métodos mencionados anteriormente se hace necesario seleccionar los valores de unos ciertos hiperparámetros para optimizar el rendimiento del modelo en cuestión. Para ello se ha utilizado una técnica conocida como *Kfold cross validation (validación cruzada)*, implementada en la librería scikit de Python. La validación cruzada [46] consiste en realizar N iteraciones, donde en cada una de ellas se selecciona una partición de los datos como datos de prueba y el resto como datos de entrenamiento, tal y como se muestra en la figura 2.5.1.

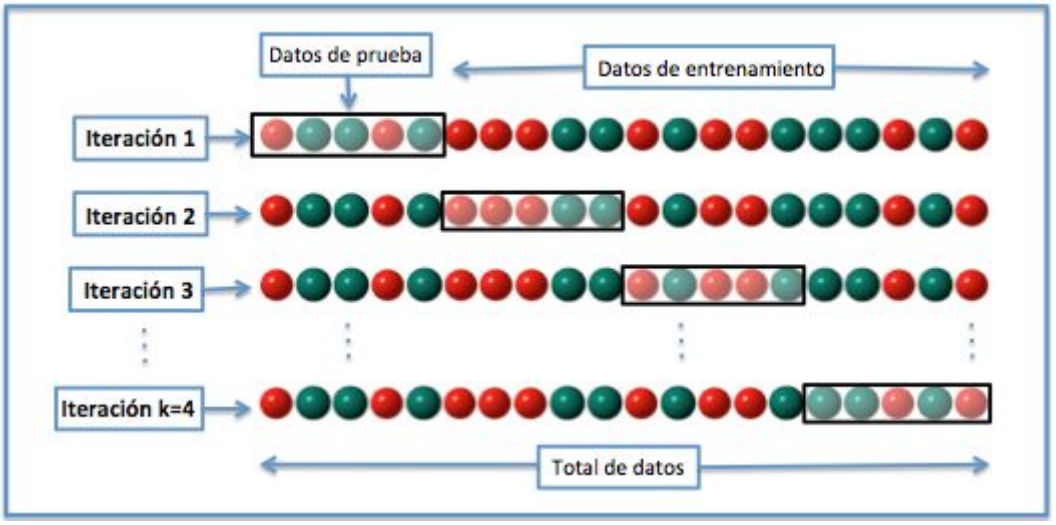


Figura 2.5.1 Ejemplo de validación cruzada con k=4 iteraciones

La validación cruzada permite asegurar que los resultados obtenidos son independientes de la partición entre datos de test y entrenamiento. La validación cruzada se puede utilizar junto con la selección de hiperparámetros de la siguiente forma:

```
for param_1 in values_param_1:
        best_error = inf
        best_params = []
        for param_n in values_param_n:
                error = 0
                for k in [1..K_fold_iterations]:
                        data_train,data_validation = split(data,k)
                        model = train(data_train,param_1, ... ,param_n)
                        prediction = predict(model,data_validation)
                        error_k = compute_error(prediction,data_validation)
                        error = error + error_k
                error = error / k
                if error < best_error:
                        Actualizar best_params con [param_1, ... ,param_n]
```

Figura 2.5.2 Algoritmo de K-fold

Al separar por datos de entrenamiento y validación también se evita que se produzca *overfitting* a la hora de seleccionar los mejores hiperparámetros.

# Materiales y métodos

Para este estudio se han utilizado los datos de ADN mitocondrial de 809 sujetos [22]. Dichos datos estudian un total de 1649 SNP (*Single Nucleotide Polymorphism)*, no obstante en los materiales adicionales del artículo [5] se proporcionan resultados para 139 SNP, de los cuales 118 son incluidos en los SNP del ADN mitocondrial.
Como fenotipo, se encontraron problemas para encontrar los datos de ADNI que se mencionan en el artículo de P.R. Ridge, ya que los diversos ficheros que se encontraban disponibles en ADNI no cubrían todos los individuos de los cuales se tenía ADN mitocondrial. Finalmente se utilizaron los datos de volumen de hipocampo izquierdo del fichero ADNIMERGE.csv para calcular su atrofia tras 2 años. A continuación se muestra un resumen estadístico de los datos utilizados.

| | Edad | Hombres/Mujeres | CN | LMCI | EMCI | AD |
|---|---|---|---|---|---|---|
| Datos íntegros | 73.24±7.07 | 446/363 | 280 | 247 | 235 | 47 |
| Datos de los sujetos disponibles | 72.27±7.13 | 252/217 | 156 | 89 | 197 | 27 |

Figura 3.1 CN:Control, LMCI:Late mild cognitive impairment, EMCI:Early mild cognitive impairment, AD:Alzheimer's disease

## 3.1 Genotipo

En este estudio el genotipado [22] se ha realizado a 809 individuos según el genoma de referencia NC_012920.1 [24] identificando, tal y como se ha mencionado anteriormente 1649 SNP, de los cuales habrá que seleccionar aquellos que aparecen en los resultados del estudio [5]. Un SNP se puede identificar por la posición en la que aparece y el nucleótido presente en el individuo en dicha posición, de forma que un SNP con nombre MitoT11900C indicaría una variación en la posición 11900, concretamente que se encuentra un nucleótido C, en lugar de T.

Los resultados del genotipado aparecen en un fichero de formato VCF (Variant Call Format) [25], en una matriz de n_individuos x n_SNP. Cada elemento i,j de esa matriz corresponderá a la información del SNP j en el individuo i. Para comprobar qué SNP hay que estudiar de cada estudio se han utilizado los siguientes campos de cada elemento:

- GT (vector): Genotipo, índice para el valor del nucleótido en el alelo correspondiente, en este caso al ser ADN mitocondrial se utiliza únicamente el índice 0. Si el valor de GT es 0 indica que el individuo no presenta ese SNP.
- ALT (vector): Alteración. Si se utiliza el valor de GT (>0) como índice se obtiene el valor del nucleótido del sujeto en la posición indicada. Para una misma posición pueden existir distintas variaciones, por lo que este campo es utilizado para asegurarse que el individuo presenta una en concreto.
- POS (entero): Posición del SNP en cuestión.

Se implementó un script en Python que comprobaba qué SNPs del estudio presentaba cada paciente en base a la información del archivo VCF, los resultados se almacenaron en un diccionario con el formato: $\{ID-USUARIO : \{NombreSNP : 0/1\}\}$ donde 1 representa la presencia de ese SNP en el sujeto.

## 3.2 Fenotipo

En este estudio se estudia el porcentaje de atrofia del hipocampo izquierdo en un periodo no inferior a 2 años, la información se obtiene del fichero UCSFFSX51_11_08_19.csv de ADNI, en donde aparecen las distintas mediciones de MRI de los sujetos en distintas fechas, a partir de lo cual se calcula el fenotipo utilizado.

# Resultados

En esta sección se van a presentar los diversos resultados obtenidos en este TFG. En el primer apartado se presentan los resultados de la comparación realizada entre la aplicación de los métodos propuestos en datos de ADN mitocondrial y en los datos de ADN nuclear utilizados por Eduardo Alonso [31]. La segunda sección compara los resultados de dichos métodos con los del artículo de P.R. Ridge [5] y el último apartado se comparan con otros artículos de la literatura.

## 4.1 Comparativa entre ADN mitocondrial y ADN nuclear

A continuación se presentan las pruebas realizadas a los métodos utilizados.

### 4.1.1 Lasso y Elastic Net

Para realizar la comparativa de resultados se implementó el estudio de frecuencias de selección mencionado en el apartado de métodos, se compararon los resultados utilizando los datos de ADNI2 del hipocampo derecho utilizados en [31]. Se realizó la selección con N = 10 repeticiones, obteniendo los siguientes resultados:

| | **Implementación propia** | | | | **Implementación de Eduardo Alonso** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Lasso** | | **Elastic net** | | **Lasso** | | **Elastic net** | |
| | **Nombre** | **Freq** | **Nombre** | **Freq** | **Nombre** | **Freq** | **Nombre** | **Freq** |
| **1** | rs769449 | 0.5 | rs769449 | 0.8 | rs769449 | 0.4 | rs769449 | 0.6 |
| **2** | rs6691117 | 0.5 | rs17014994 | 0.8 | rs2242601 | 0.4 | rs2242601 | 0.5 |
| **3** | rs6458566 | 0.5 | rs6967117 | 0.7 | rs6458566 | 0.3 | rs2404529 | 0.4 |
| **4** | rs7533408 | 0.4 | rs2565050 | 0.8 | rs12616704 | 0.3 | rs6458566 | 0.4 |
| **5** | rs4726618 | 0.4 | rs17057441 | 0.7 | rs9314349 | 0.3 | rs12616704 | 0.3 |

Tabla 4.1.1.1 5 primeros SNP ordenados por frecuencias

Se puede observar que en todos los casos el SNP rs769449, del gen APOE es el que presenta una mayor frecuencia. El gen APOE aparece en la literatura como un marcador de riesgo de la enfermedad [5][31][32]. También hay coincidencias entre otros SNP como el rs6458566 en el caso de Lasso. Es importante mencionar que existen numerosos empates, y pequeñas variaciones en la frecuencia que afectan al ordenamiento de los SNP. Esto se podría atenuar de cierta forma incrementando N sustancialmente, aunque requeriría un gran esfuerzo computacional. Por ejemplo el SNP rs2242601 presenta una frecuencia de 0.4 en la implementación propia de Lasso y de 0.6 en elastic net, mientras que el SNP rs12616704 presenta una frecuencia en Lasso de 0.1 y en Elastic net de 0.5. Se puede observar que pese a ser de distintas implementaciones se obtienen unos resultados similares, lo que nos permite comprobar en un entorno más controlado que el de ADN mitocondrial que los métodos de entrenamiento de los modelos funcionan correctamente.

### 4.1.2 SVM

Para comprobar el funcionamiento del modelo *"stepwise"* mencionado en el apartado 2.6.4 este se ha utilizado con los mismos datos mencionados en el apartado anterior y tomando como guía el ránking de genes utilizados en [31], de forma que los SNP pertenecientes a los genes que tengan una posición más alta en el ránking serán añadidos antes. Se han obtenido los siguientes resultados:

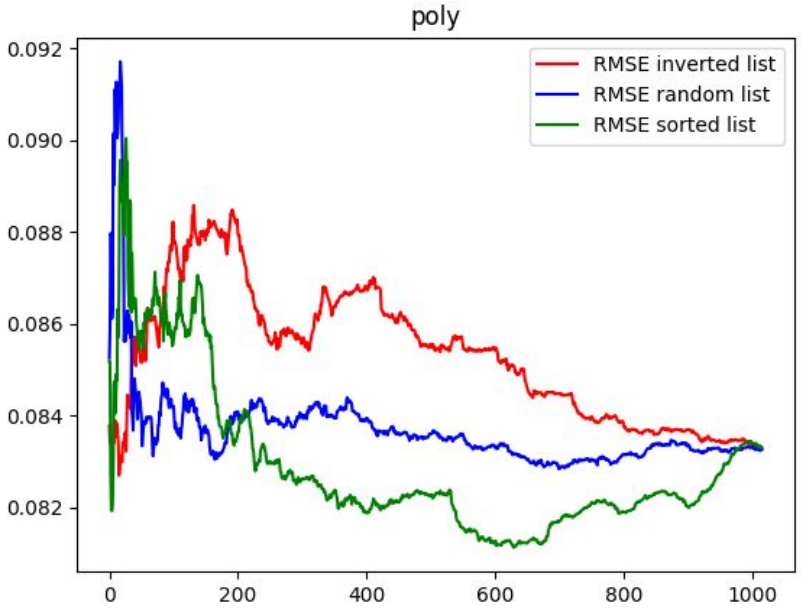


Figura 4.1.2.1 Evolución del RMSE según el número de SNPs añadidos utilizando el kernel polinomial (poly)

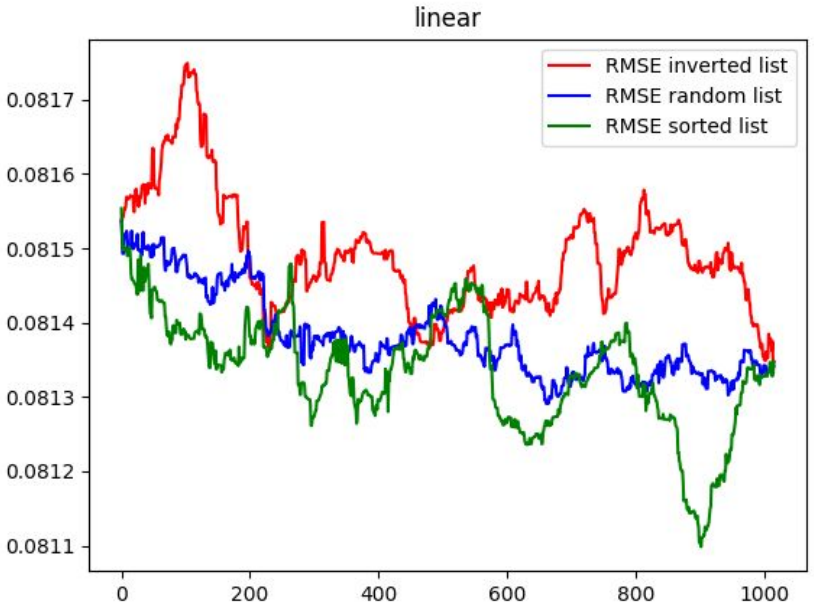


Figura 4.1.2.2 Evolución del RMSE según el número de SNPs añadidos utilizando el kernel lineal

Como se puede observar, el modelo tiene mayor error añadiendo primero los SNP del final del ránking (los que se supone que tienen menor relación con el Alzheimer), además añadiendo los SNP según el orden de dicho ránking se alcanza un mínimo a partir del cual añadir más SNP supone añadir error al modelo.

Para comprobar que las fluctuaciones en el error se debían a la naturaleza de los datos añadidos y no a otros factores como el número de características se llevó a cabo una prueba con el kernel lineal identificando las regiones del eje x que correspondían a cada gen.

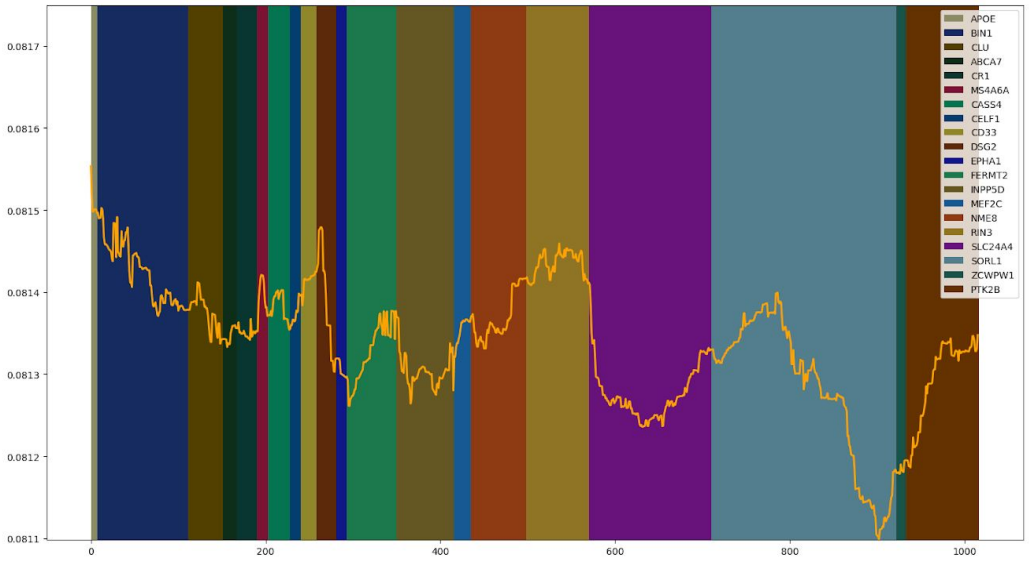


Figura 4.1.2.3 Evolución del RMSE añadiendo los SNP ordenadamente según el ránking

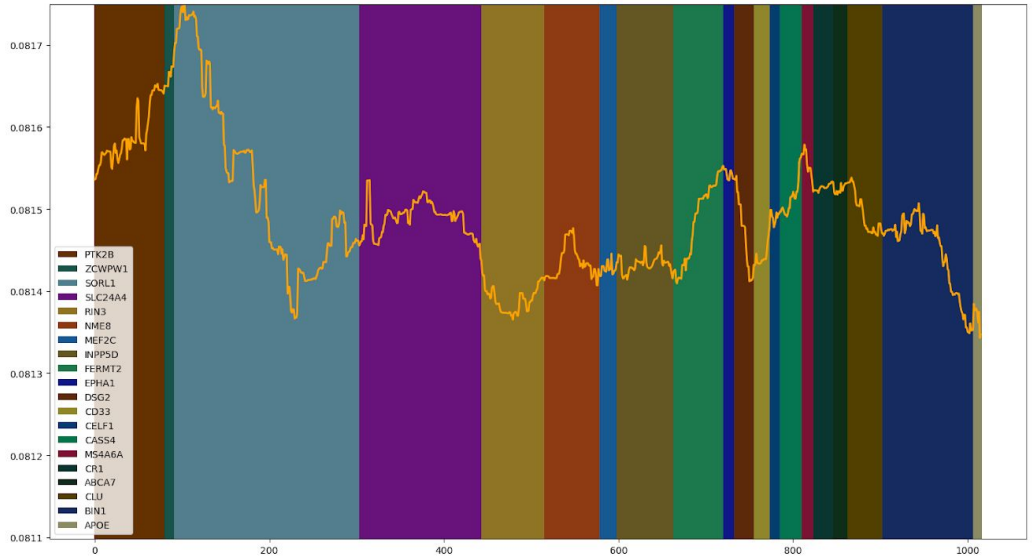


FIgura 4.1.2.4 Evolución del RMSE añadiendo los SNP en orden inverso

Cada región de color pertenece a un gen en concreto. Se puede observar que, independientemente del orden en el que se realicen las inclusiones al modelo, hay tendencias que se repiten. Por ejemplo, en ambos casos los SNP del gen BIN1 (Azul oscuro del principio en la lista ordenada) producen una bajada del error, o los SNP situados entre los genes SLC24A4 (morado) y RIN3 (amarillo) también reducen el error, entre otros. Por lo cual se ha comprobado que, efectivamente el método propuesto de stepwise se puede utilizar para determinar qué SNPs tienen una mayor relación con el fenotipo estudiado.

## 4.2 Comparativa con los resultados de P.G. Ridge

En esta sección, una vez realizadas las pruebas para comprobar el buen funcionamiento de los métodos propuestos con ADN nuclear, se comparan los resultados obtenidos con los que aparecen en el artículo de P.G. Ridge [5] con ADN mitocondrial.

### 4.2.1 Lasso y Elastic Net

Para comparar los resultados con los modelos de Lasso y Elastic net se ha comparado la posición en el ránking según su coeficiente medio tras 10 iteraciones del algoritmo propuesto en el apartado 2.6.2.

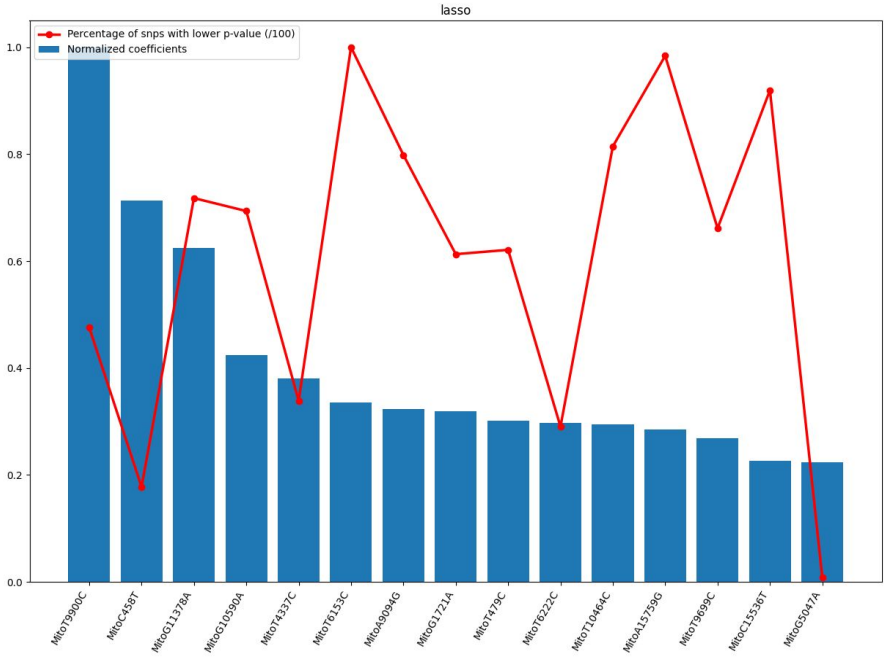


Figura 4.2.1.1 Comparativa de los coeficientes medios y la posición de los SNP en el ránking de P.R. Ridge

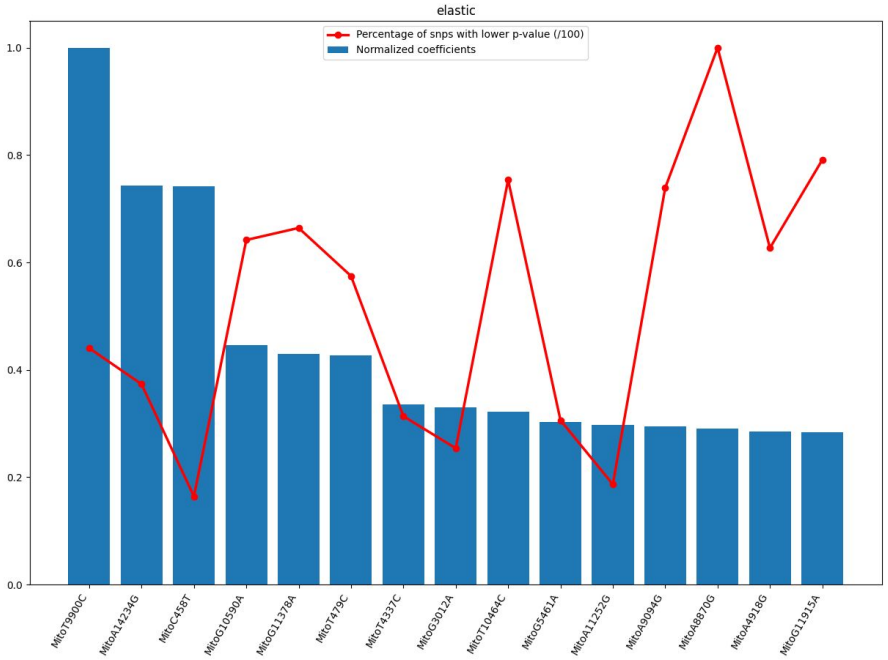


Figura 4.2.1.2 Comparativa entre los coeficientes medios y el ránking en la posición de Perry G. Ridge

Los resultados obtenidos no concuerdan con los que presenta P.G. Ridge en su artículo. No obstante, se puede observar que en ambos casos hay SNPs que aparecen elegidos por ambos modelos, como puede ser el caso de MitoT9900C o MitoC458T.

### 4.2.2 SVM

Realizando el método de stepwise con los datos del ADN mitocondrial se obtienen los siguientes resultados:

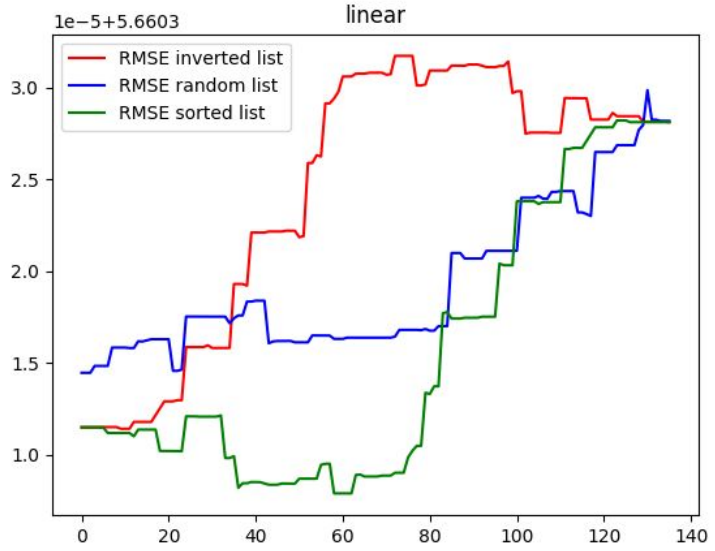


Figura 4.2.2.1 Evolución del RMSE con el kernel lineal

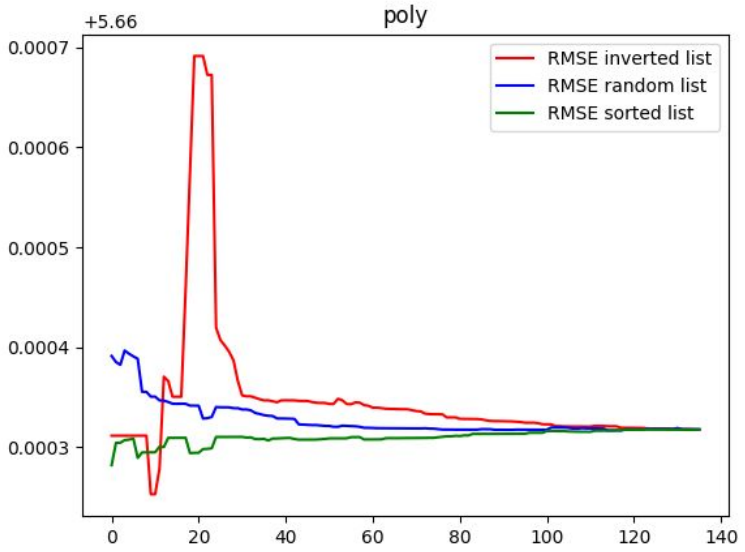


Figura 4.2.2.2 Evolución del RMSE con kernel polinómico

Comparándolo con los resultados en el ADN nuclear del apartado 4.1.2 se puede observar un comportamiento similar, especialmente en el caso del kernel lineal. En el caso del kernel polinómico, aparece un comportamiento muy irregular al inicio. Los resultados del caso lineal muestran además el mismo comportamiento esperado que en el caso del ADN nuclear, obteniendo el menor valor de RMSE al incluir los primeros 60 SNP del artículo de P.G. Ridge [5].

## 4.3 Comparativa con otros artículos

Además se ha buscado en la literatura otros resultados con el fin de averiguar si los SNP que se han elegido en este estudio han sido relacionados con algún tipo de proceso biológico que pudiese estar relacionado con la enfermedad de Alzheimer. De dichos artículos [32], [33] se han extraído varios SNP: MitoT14179, MitoA11468G, MitoA12309G y MitoG12373 y se han observado los resultados obtenidos con ellos en este estudio.

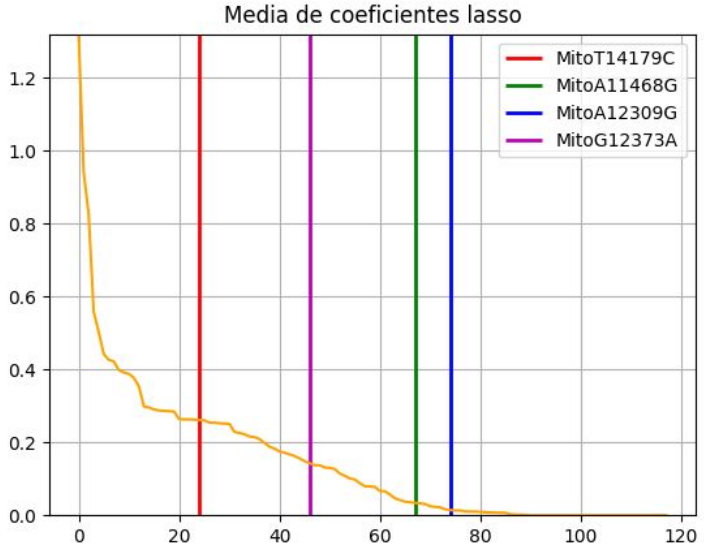


Figura 4.3.1 Coeficientes de Lasso ordenados y posición de los SNP

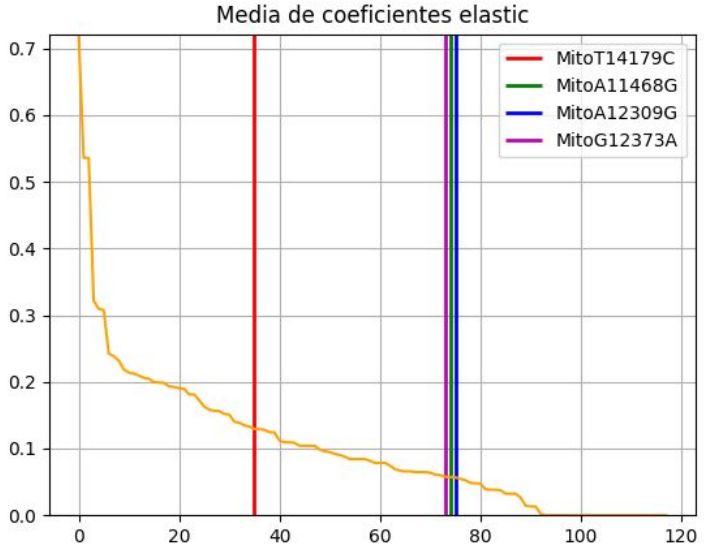


Figura 4.3.2 Coeficientes de Elastic net ordenados y posición de los SNP

En las dos figuras anteriores se muestra los coeficientes de los modelos Elastic net y Lasso ordenados de mayor a menor peso y la posición de los SNP encontrados. En ambos casos el SNP MitoT14179C se encuentra en el top 30 de los SNP, mientras que los restantes ocuparían posiciones más altas.

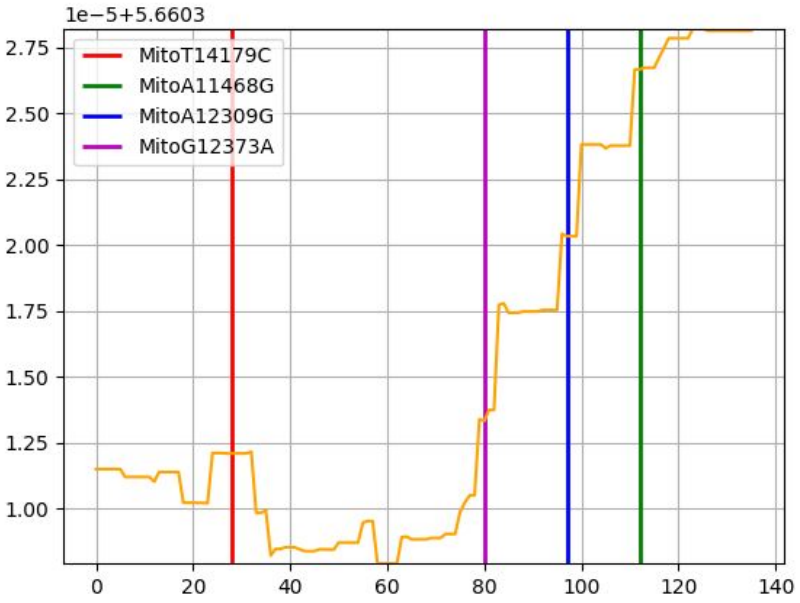


Figura 4.3.3 Evolución del error RMSE y posición de los SNP

En la figura anterior se muestra la evolución del error tras realizar el algoritmo stepwise y la posición de los SNP de los artículos. Una vez más el SNP MitoT14179C está en el top 30, remarcando que en este último caso el orden de los SNP es el de la lista del artículo de P.G. Ridge. Respecto al comportamiento del error, se aprecia que la inclusión de los SNP señalados no supone un descenso en él ni tampoco un aumento.

## 4.4 Group Lasso

Los resultados en Group Lasso indican una fuerte penalización en el número de elementos de los grupos, lo que implica que el agrupamiento de los SNP en este caso no mejora los resultados, por lo que se descartaron los resultados obtenidos por este método.

## 4.5 Estudio de haplogrupos

Se han estudiado los haplotipos de los sujetos que presentan los SNP MitoT9900C y MitoC458T, obteniendo que el haplogrupo predominante en el caso del SNP MitoT9900C es el haplogrupo T, mientras que en el caso de MitoC458T es el haplogrupo H5 y H5/A. Como se puede observar en las figuras 4.2.1.1 y 4.2.1.2 el SNP MitoC458T es el que tiene más importancia según los resultados propuestos por P.G. Ridge. Precisamente en otro artículo [46] en el que participa este mismo autor se presenta una tabla donde aparecen sintetizados otros haplogrupos que han aparecido en la literatura:

| Haplogroup | Dataset | Effect | Ethnicity | # cases / controls |
|---|---|---|---|---|
| B4C1 (Takasaki 2009) [47] | Selected SNPs | Risk | Japanese | 96 / 384 |
| G2A (Takasaki 2009) [47] | Selected SNPs | Risk | Japanese | 96 / 384 |
| HV (Maruszak et al. 2009) [48] | Haplogroups, SNPs | Risk | Polish | 222 / 252 |
| H (Fesahat et al. 2007) [49] | HVS-I sequence | Risk | Iranian | 30 / 100 |
| H5 / H5A (Santoro et al. 2010) [50] | D-loop sequence, restriction analysis | Risk | Italian | 936 / 776 |
| H6A1A / H6A1B (Ridge et al. 2012) [51] | Full mtDNA sequences | Protective | Caucasian | 101 / 632 |
| K (Carrieri et al. 2001) [52] | Haplogroups | Protective | Italian | N/A* |
| N9B1 (Takasaki 2009) [47] | Selected SNPs | Risk | Japanese | 96 / 384 |
| U (van der Walt et al. 2004; Fesahat et al. 2007) [53][49] | HVS-I sequence, 10 SNPs | Risk | Iranian, Caucasian | 30 / 100, 989 / 328** |
| U (Carrieri et al. 2001; van der Walt et al. 2004)[52] | Haplogroups, 10 SNPs | Protective | Italian, Caucasian | N/A*, 989 / 328** |
| UK (Lakatos et al. 2010) [33] | 138 SNPs | Risk | Caucasian | 170 / 188 |
| None (Zsurka et al. 1998) [54] | 4 SNPs | None | Unknown | 70 / 80 |
| None (Chinnery et al. 2000) [55] | European Haplogroups | None | Unknown | 185 / 179 |
| None (Pyle et al. 2005) [56] | U, K, J, and T haplogroups | None | English | 185 / 447 |
| None (Mancuso et al. 2007) [57] | European Haplogroups | None | Tuscan | 209 / 191 |

| | | | | |
|---|---|---|---|---|
| None (Kruger et al. 2010) [58] | Haplogroups | None | Finnish | 128 / 99*** |
| None (Hudson et al. 2012) [59] | 138 SNPs | None | Caucasian | 3250 / 1221 |

Figura 4.5.1 Tabla con los resultados de los distintos artículos científicos
* Los autores mostraron que los haplogrupos U y K neutralizan el riesgo del alelo e4 de APOE
** Los autores demostraron un riesgo mayor para hombres con haplogrupo U, y menor para mujeres con haplogrupo U
*** Se trataba de casos Alzheimer con inicio temprano

En la tabla, se puede observar que los haplogrupos H5 y H5A aparecen como un indicador de riesgo en Santoro et al. 2010 [50] lo que puede tener relación con el hallazgo del SNP MitoC458T como relevante.

# Conclusión

La enfermedad de Alzheimer es una enfermedad compleja y desconocida en muchos aspectos a pesar de los estudios actuales. Todavía no se conocen con exactitud cuáles son los factores que favorecen su desarrollo y que podrían ser fundamentales para un diagnóstico precoz. Este tipo de diagnóstico es fundamental para la aplicación de los tratamientos paliativos que existen y que permiten retrasar la aparición de la demencia, y podría ser la clave para encontrar un tratamiento para curar esta enfermedad.
En este TFG se ha abordado una línea de investigación que trata de relacionar los datos genéticos del ADN mitocondrial de los sujetos con los biomarcadores anatómicos de referencia para la enfermedad de Alzheimer. Esta línea de investigación ha sido iniciada hace pocos años por unos pocos grupos de investigación en todo el mundo. En general, la genética es un campo de la ciencia en el que existen muchas incógnitas, especialmente en el papel que ésta presenta en el desarrollo de ciertas enfermedades multifactoriales. Desde un punto de vista de ciencia de datos, los datos genéticos suponen una dificultad añadida: además de todos los esfuerzos previos que se requieren para su utilización (secuenciación, alineamiento…) son datos que representan una realidad biológica muy compleja. Recordemos, por ejemplo, que el ADN es el encargado de codificar la información para sintetizar proteínas y compuestos biológicos que están involucrados en millones de procesos biológicos. Es por esto que el comportamiento de la atrofia del hipocampo izquierdo (que es el fenotipo que se ha analizado en este trabajo respecto a las mutaciones de su ADN mitocondrial) puede estar sujeto a otros factores que dependen del sujeto, como por ejemplo, la combinación de otros SNPs presentes, factores genéticos del ADN nuclear o incluso factores ambientales y del modo de vida, el cual se ha demostrado que puede influir en la evolución de la enfermedad. Por esto, la reproducibilidad de los resultados de un estudio es un factor muy importante a la hora de determinar si un SNP puede estar relacionado con la enfermedad.
Como objetivo inicial de este TFG se intentaron reproducir los resultados obtenidos por P.R. Ridge et al. en [5]. No obstante, se han encontrado numerosas dificultades que han impedido la reproducción de dichos resultados con la misma aproximación al problema que se hizo en [5]. La primera y más notable ha sido la falta de disponibilidad del programa utilizado en dicho artículo para establecer las relaciones entre genotipo y fenotipo, lo que nos ha llevado a abordar el problema utilizando una metodología diferente. En segundo lugar, los datos encontrados en ADNI no coinciden con los utilizados en el artículo, debido a que estos datos están sujetos a posibles modificaciones por los organismos que los gestionan en ADNI.

Así, en este TFG se ha implementado una metodología basada en métodos de selección de características en aprendizaje automático, la cual ha sido probada inicialmente con ADN nuclear para comprobar su validez y posteriormente ser utilizada con ADN mitocondrial. Dichos métodos obtienen resultados distintos que los propuestos por P.R. Ridge [5]. A diferencia de los estudios con ADN nuclear, donde los SNP del gen APoE son consistentemente seleccionados como relevantes para la enfermedad, en ADN mitocondrial todavía no se ha encontrado ningún SNP inequívocamente relacionado con la enfermedad. Este hecho, la complejidad de los datos manejados y las diferencias en la selección de los sujetos respecto al estudio de P.R. Ridge, justifican no haber podido reproducir los resultados del estudio original. No obstante, hay que resaltar que en nuestro estudio dos SNP han sido seleccionados consistentemente tanto por el modelo de Lasso como Elastic-net (MitoT9900C y MitoC458T). Esto, unido con el estudio de haplotipos realizado que vinculan al SNP MitoC458T con los haplogrupos H5 /H5A, los cuales han sido identificados como posibles marcadores de riesgo de la enfermedad en [50], podría ser un indicativo de que estos SNP tienen algún tipo de relación con la enfermedad lo cual debería ser corroborado por estudios de carácter biológico. Además los resultados del modelo SVM junto con la comparativa con los SNP obtenidos de otros estudios parecen indicar que efectivamente, el ADN mitocondrial podría afectar al desarrollo de la enfermedad de Alzheimer.

En definitiva, el Alzheimer es una enfermedad cuyas causas se desconocen, aunque se cree y existen evidencias de que tiene un fuerte componente genético. En el campo de la genética existen aún muchas incógnitas en cómo ésta puede afectar al desarrollo de muchas enfermedades y el Alzheimer no es una excepción. Pese a ello, se han obtenido resultados que indican una posible relación entre la enfermedad del Alzheimer y ciertas variaciones en el ADN nuclear y mitocondrial. En este TFG se ha seguido la dirección señalada por el reciente estado del arte [45]:

*There is significant evidence for the role of mitochondria in AD risk. Studies of the contribution of mitochondrial genetic variation to AD risk remain inconclusive due to small sample sizes, limited genetic data collection, and inadequate approaches to association analysis.*

Creemos que los resultados obtenidos en este TFG constituyen una pequeña pero importante contribución a este estado del arte, que deja abierta la puerta al diseño de algoritmos que permitan encontrar o refutar la relación entre el ADN mitocondrial y el desarrollo de la enfermedad de Alzheimer. Así, como trabajo futuro se podría abordar el problema utilizando técnicas de deep learning para la selección de los SNP más relevantes. También se podría llegar a utilizar información genética nuclear y mitocondrial para obtener una visión global de los genes, tanto del ADN nuclear como del ADN mitocondrial que influyen en la enfermedad y si existe alguna relación entre ellos. En esta última línea de investigación sería además posible la utilización de otras técnicas como la estudiada en el Anexo 1.

# Anexos

# Anexo 0: Desarrollo de este TFG


Figura A.0 Diagrama de Gannt con el tiempo dedicado a las distintas tareas

En la figura anterior se muestra un diagrama de Gannt con el tiempo aproximado de realización de cada una de las tareas. Se inició leyendo la información relativa a la tesis utilizada como referencia, así como de los métodos propuestos en ella. Tras decidir el cambio de metodología anteriormente comentado se tuvo que otro proceso de investigación acerca de la nueva metodología a utilizar,continuada de la recolección y preprocesamiento de los datos a utilizar. Finalmente se llevó a cabo la implementación de los modelos elegidos, una vez se obtuvieron los resultados de dichos modelos se procedió a la búsqueda en la literatura acerca de estos resultados para, finalmente, realizar la redacción de la memoria.

Cabe destacar que la situación acaecida este año 2020 ha propiciado una dilatación general en la duración de las tareas.

# Anexo 1: Enriquecimiento genético

Durante la realización de este TFG se han tratado de realizar diversas técnicas de tratamiento de datos genéticos que, debido a la naturaleza de los datos utilizados no han podido llevarse a cabo. Una de esas técnicas es denominada enriquecimiento genético [14] utilizando la herramienta online Enrichr [15] que permite, dada una lista de genes, obtener las funciones biológicas que se ven más relacionadas con esos genes. El enriquecimiento genético permitiría agrupar a los SNP por funciones biológicas además de por el gen al que pertenecen. No obstante, se descubrió que esta técnica necesita un gran número de genes que es muy limitado en el ADN mitocondrial lo que nos llevó a replantear esta línea de investigación.

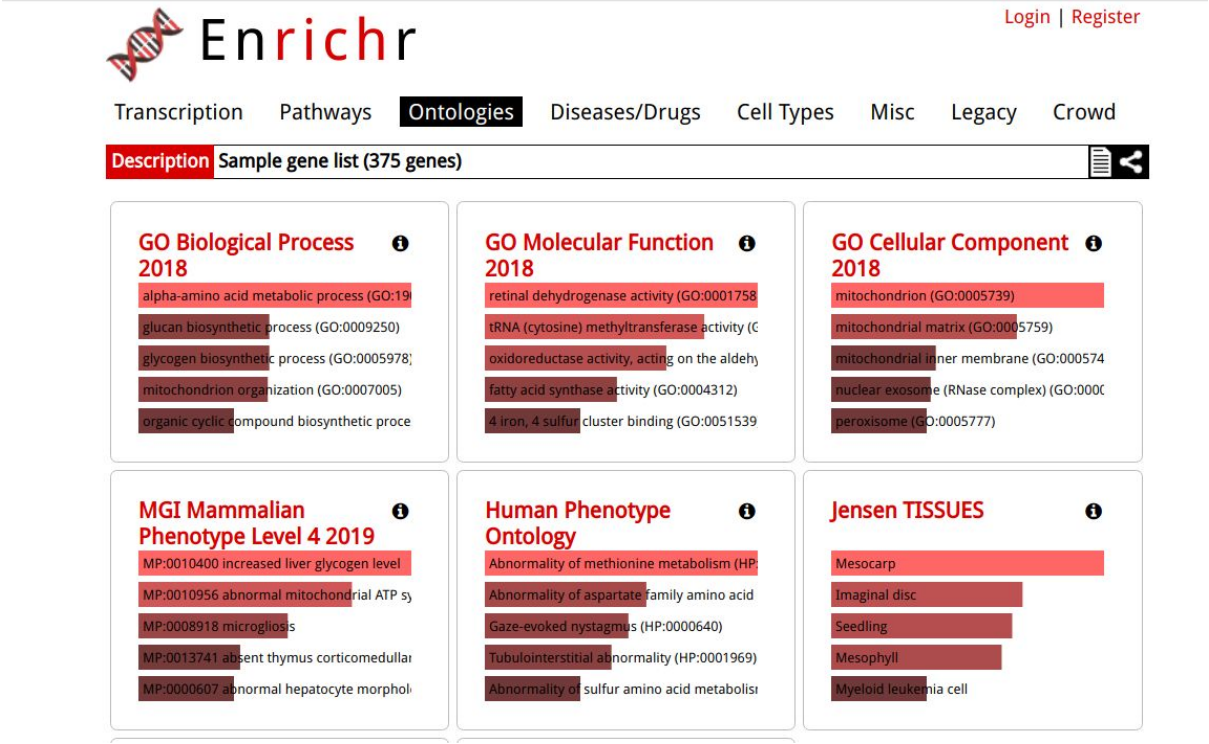


Figura A.1 Captura de los resultados obtenidos con Enrichr utilizando un set de genes de ejemplo

# Anexo 2: Resultado con otros fenotipos

En este anexo se presentan los resultados obtenidos para otros fenotipos propuestos por P.G. Ridge: el grosor del polo temporal y el volumen cerebral. Estos resultados han sido descartados del estudio principal ya que se ha considerado que no aportan nada a éste. Además como se menciona en el trabajo el hipocampo es una de las regiones que antes sufren las consecuencias del Alzheimer, lo que también llevó a centrarnos en esta región del cerebro.

## Anexo 2.1 Volumen cerebral

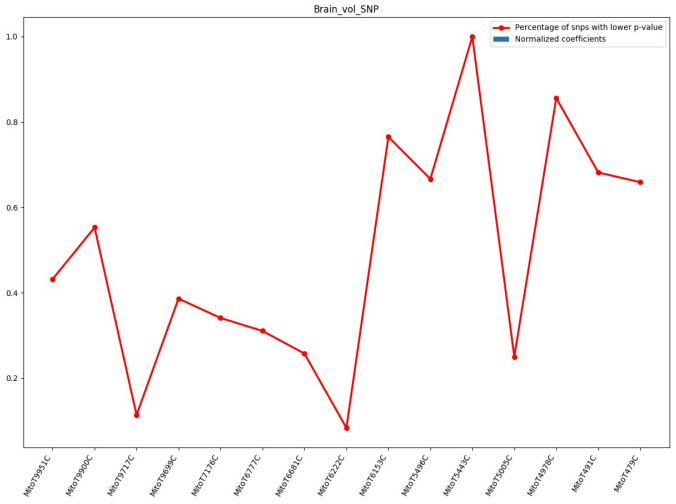

Figura A.2 comparativa de coeficientes de Elastic net con los resultados de P.G Ridge
Se puede observar que en este caso el modelo no ha seleccionado ningún SNP

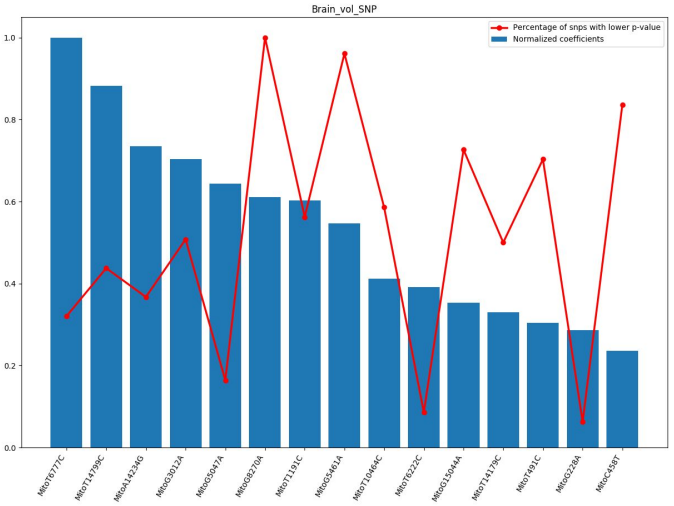


Figura A.3 comparativa de coeficientes de Lasso con los resultados de P.G Ridge

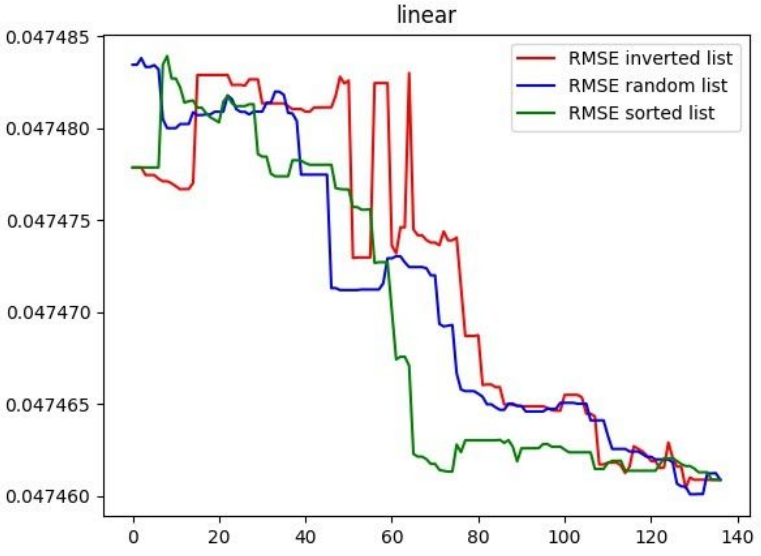


Figura A.4 evolución  del error RMSE según los SNP añadidos con kernel lineal

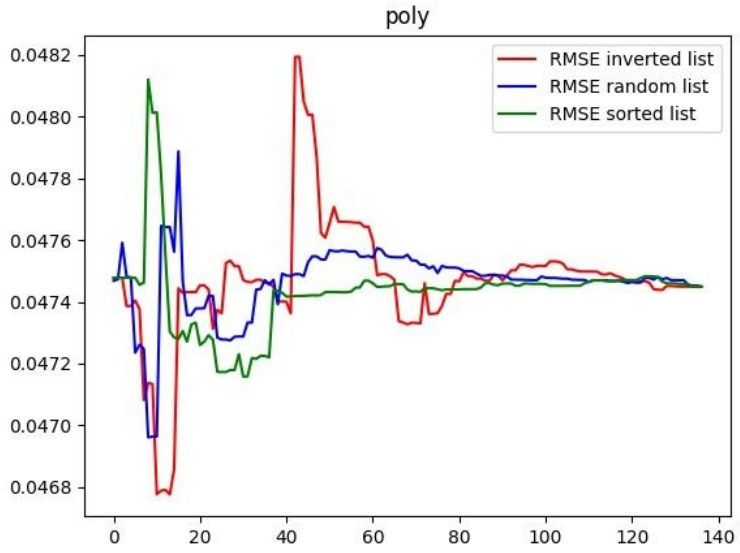


Figura A.5 evolución del error RMSE según los SNP añadidos con kernel polinómico

# Anexo 2.2 Grosor del polo temporal

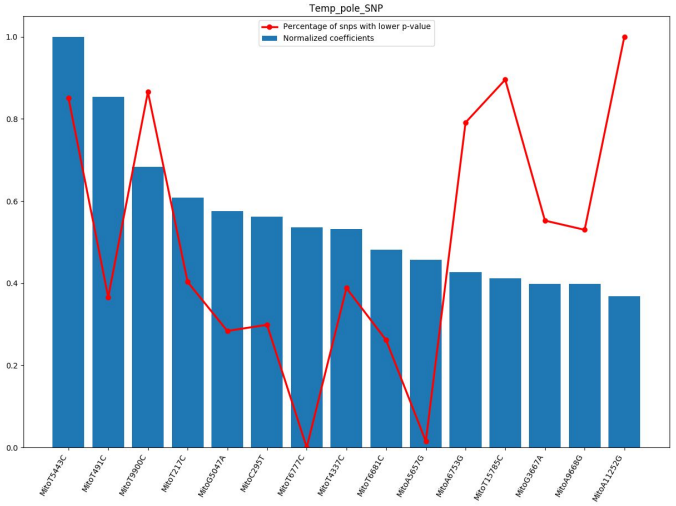


Figura A.6 Coeficientes Lasso en comparación con los resultados de P.G. Ridge

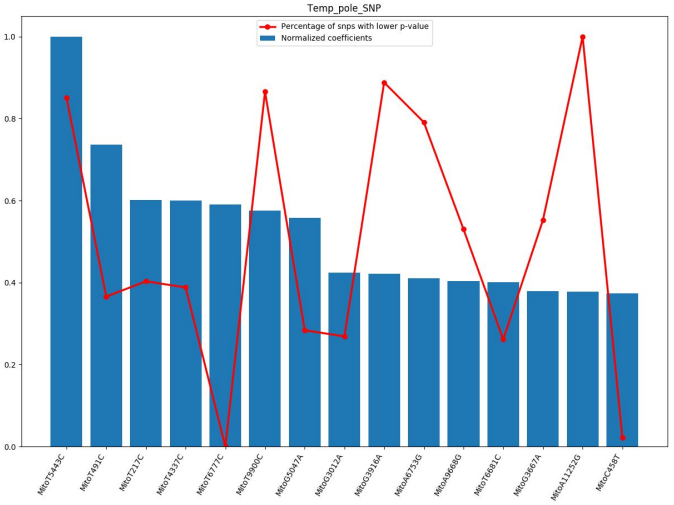


Figura A.7 Coeficientes Elastic net en comparación con los resultados de P.G. Ridge

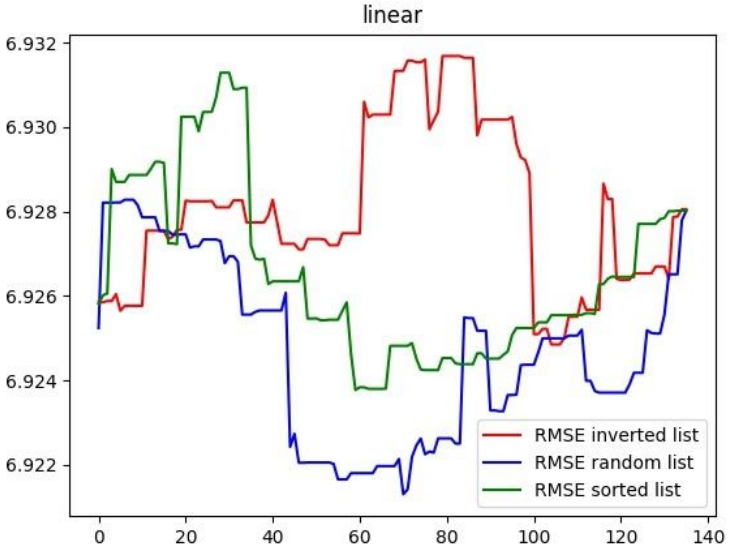


Figura A.8 evolución del error RMSE según los SNP añadidos con kernel lineal

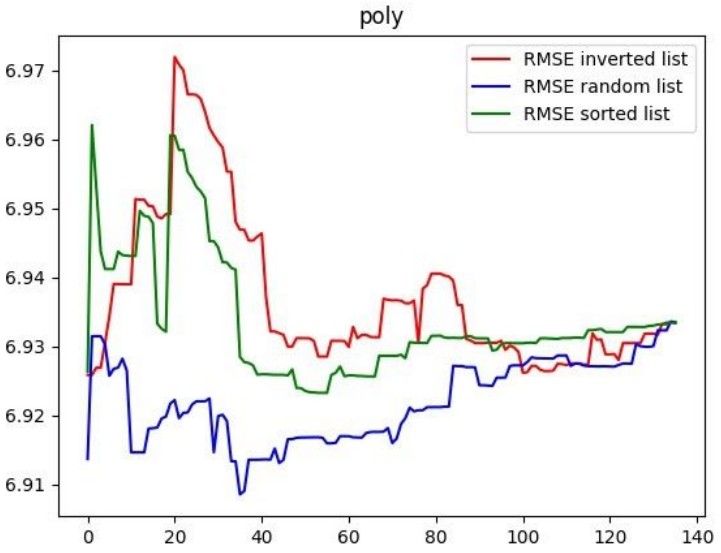


Figura A.9 evolución del error RMSE según los SNP añadidos con kernel polinómico

# Bibliografía

**[1]**
Alzheimer disease: Alzheimer Disease Dis Mon.(2010)
Rudy J. Castellani ,Raj K. Rolston, and Mark A. Smith
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2941917/
**[2]**
Reduced gray matter volume in normal adults with a maternal family history of Alzheimer disease Neurology (2010)
R. A. Honea, R. H. Swerdlow, E. D. Vidoni, J. Goodwin, J. M. Burns
https://n.neurology.org/content/74/2/113.short
**[3]**
Healthy lifestyle and the risk of Alzheimer dementia Neurology (2020)
Klodian Dhana, Denis A. Evans, Kumar B. Rajan, David A. Bennett, Martha C. Morri
https://n.neurology.org/content/95/4/e374.abstract
**[4]**
Effect of lifestyle activities on alzheimer disease biomarkers and cognition Annals of Neurology (2012)
Prashanthi Vemuri PhD et al.
https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.23665
**[5]Perry**
Mitochondrial Genetics of Alzheimer's Disease and
Aging current genetic medicine reports (2013)
Perry Gene Ridge
**[6]**
Regression Shrinkage and Selection via the Lasso Journal of the Royal Statistical Society (1996)
Robert Tibshirani
https://www.jstor.org/stable/2346178?seq=1
**[7]**
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

**[8]**
Regularization and variable selection via the elastic net Journal of the Royal Statistical Society (2005)
Hui Zou Trevor Hastie
https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1467-9868.2005.00503.x
**[9]**
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html#sklearn.linear_model.ElasticNet
**[10]**
A Sparse-Group Lasso Journal of Computational and Graphical Statistics(2013)
Noah Simon, Jerome Friedman, Trevor Hastie,
and Rob Tibshirani
https://www.tandfonline.com/doi/full/10.1080/10618600.2012.681250
**[11]**
Model selection and estimation in regression with
grouped variables Journal of the Royal Statistical Society Series B (Statistical Methodology) (2006)
Ming Yuan,Yi Lin
https://www.researchgate.net/publication/4993325_Model_Selection_and_Estimation_in_Regression_With_Grouped_Variables
**[12]**
https://group-lasso.readthedocs.io/en/latest/api_reference.html

**[13]**
https://en.wikipedia.org/wiki/Mitochondrial_DNA
**[14]**
https://en.wikipedia.org/wiki/Gene_set_enrichment_analysis
**[15]**
https://maayanlab.cloud/Enrichr/
**[16]**
https://towardsdatascience.com/truly-understanding-the-kernel-trick-1aeb11560769
**[17]**
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
**[18]**
https://en.wikipedia.org/wiki/Stepwise_regression
**[19]**
TCS: a computer program to estimate gene genealogies.Molecular Ecology Notes (2000)
Clement M, Posada D, Crandall KA
https://www.researchgate.net/publication/248832683_TCS_A_computer_program_to_estimate_gene_genealogies
**[20]**
Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. PubMed (2005)
Templeton AR, Maxwell T, Posada D, Stengard JH, Boerwinkle E, et al.
https://www.researchgate.net/publication/8344190_Tree_Scanning_A_Method_for_Using_Haplotype_Trees_in_PhenotypeGenotype_Association_Studies
**[21]**
TreeScan: a bioinformatic application to search for genotype/phenotype associations using haplotype trees. Bioinformatics (Oxford, England)(2005)
Posada D, Maxwell TJ, Templeton AR

https://europepmc.org/article/med/15681571
**[22]**
Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping Alzheimer's & Dementia (2018)
Ridge et al
http://adni.loni.usc.edu/adni-publications/Assembly%20of%20809%20whole%20mitochondrial%20genomes%20with%20clinical,%20imaging,%20and%20fluid%20biomarker%20phenotyping.pdf
**[23]**
https://es.wikipedia.org/wiki/%C3%81cido_ribonucleico
**[24]**
https://www.ncbi.nlm.nih.gov/nuccore/251831106
**[25]**
https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/
**[26]**
https://es.wikipedia.org/wiki/Fenotipo
**[27]**
https://es.wikipedia.org/wiki/Hipocampo_(anatom%C3%ADa)#Anatom%C3%ADa
**[28]**
Loss of Recent Memory After Bilateral Hippocampal Lesions J Neurol Neurosurg Psychiatry. (1957)
Scoville, WB; Milner B
https://web.archive.org/web/20080504040528/http://neuro.psychiatryonline.org/cgi/content/full/12/1/103
**[29]**
Relationship between hippocampal atrophy and neuropathology markers: A 7T MRI validation study of the EADC-ADNI Harmonized Hippocampal Segmentation Protocol Alzheimers Dement(2015)
Liana G. Apostolova et al
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4348340/
**[30]**
Core candidate neurochemical and imaging biomarkers of Alzheimer's disease Alzheimers Dement (2008)
Harald Hampel et al
https://pubmed.ncbi.nlm.nih.gov/18631949/
**[31]**
Un estudio de asociación genómica basado en aprendizaje automático para la caracterización de la enfermedad de Alzheimer (Trabajo de Final de Grado, Universidad de Zaragoza, 2020)
Eduardo Alonso Monge

**[32]**
Mitochondria and Alzheimer’s Disease: the Role of Mitochondrial Genetic Variation Curr Genet Med Rep (2018)
Perry G. Ridge and John S. K. Kauwe
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5842281/
**[33]**
Association between mitochondrial DNA variations and Alzheimer's Disease in the ADNI cohort Neurobiol Aging (2010)
Anita Lakatos et al.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2918801/
**[34]**

Alzheimer's disease: the amyloid cascade hypothesis Science (1992)
Hardy, JA, GA Higgins.
https://pubmed.ncbi.nlm.nih.gov/1566067/
**[35]**
Alzheimer's disease. N Engl J Med (2010)
Querfurth, HW, FM LaFerla.
https://pubmed.ncbi.nlm.nih.gov/20107219/
**[36]**
https://es.wikipedia.org/wiki/Placa_senil
**[37]**
The pathogenesis of Alzheimer's disease: a reevaluation of the "amyloid cascade hypothesis"
Int J Alzheimers Dis (2011)
Armstrong, RA.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3038555/
**[38]**
Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. Nature (1991)
Goate, A, MC Chartier-Harlin, M Mullan, et al.
https://pubmed.ncbi.nlm.nih.gov/1671712/
**[39]**
Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease Nature (1995)
Sherrington, R, EI Rogaev, Y Liang, et al.
https://pubmed.ncbi.nlm.nih.gov/7596406/
**[40]**
Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science (1995)
Levy-Lahad, E, W Wasco, P Poorkaj, et al.
https://pubmed.ncbi.nlm.nih.gov/7638622/
**[41]**
Fibrillar amyloid-beta burden in cognitively normal people at 3 levels of genetic risk for Alzheimer's disease. Proc Natl Acad Sci U S A (2009)
Reiman, EM, K Chen, X Liu, et al.
https://pubmed.ncbi.nlm.nih.gov/19346482/

**[42]**
Polymorphisms in the human apolipoprotein-J/clusterin gene: ethnic variation and distribution in Alzheimer's disease. Human Genetics (1996)
Tycko, B, L Feng, L Nguyen, et al.
https://link.springer.com/article/10.1007/s004390050234
**[43]**
Genetic analysis of Alzheimer's disease in the Uppsala
Longitudinal Study of Adult Men. Dement Geriatr Cogn Disord (2009)
Giedraitis, V, L Kilander, M Degerman-Gunnarsson, J Sundelof, T Axelsson, AC Syvanen, L Lannfelt, A Glaser.
https://pubmed.ncbi.nlm.nih.gov/19141999/
**[44]**
Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. (2009)
Harold, D, R Abraham, P Hollingworth, et al.

https://pubmed.ncbi.nlm.nih.gov/19734902/
**[45]**
Mitochondria and Alzheimer's Disease: the Role of Mitochondrial Genetic Variation Nat Genet (2018)
Perry G. Ridge, John S. K. Kauwe
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5842281/
**[46]**
https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada
**[47]**
Mitochondrial haplogroups associated with Japanese Alzheimer's patients. JBioenerg Biomembr (2009)
Takasaki, S
https://pubmed.ncbi.nlm.nih.gov/19795196/
**[48]**
Mitochondrial haplogroup H and Alzheimer's disease--is there a connection? Neurobiol Aging (2009)
Maruszak, A, JA Canter, M Styczynska, C Zekanowski, M Barcikowska.
https://pubmed.ncbi.nlm.nih.gov/18308428/
**[49]**
Do haplogroups H and U act to increase the penetrance of Alzheimer's disease? Cell Mol Neurobiol (2007)
Fesahat, F, M Houshmand, MS Panahi, K Gharagozli, F Mirzajani.
https://pubmed.ncbi.nlm.nih.gov/17186363/
**[50]**
Evidence for sub-haplogroup h5 of mitochondrial DNA as a risk factor for late onset Alzheimer's disease. PLoS One (2010)
Santoro, A, V Balbi, E Balducci, et al.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012037
**[51]**
Mitochondrial genomic analysis of late onset Alzheimer's disease reveals protective haplogroups H6A1A/H6A1B: the Cache County Study on Memory in Aging. PLoS (2012)
Ridge, P, T Maxwell, C Corcoran, M Norton, J Tschanz, E O'Brien, R Kerber, R Cawthon, R Munger, K JSK.
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0045134
**[52]**
Mitochondrial DNA haplogroups and APOE4 allele are non-independent variables in sporadic Alzheimer's disease. Hum Genet (2001)
Carrieri, G, M Bonafe, M De Luca, et al.
https://link.springer.com/article/10.1007/s004390100463
**[53]**
Analysis of European mitochondrial haplogroups with Alzheimer disease risk. Neurosci Lett (2004)
van der Walt, JM, YA Dementieva, ER Martin, et al.
https://pubmed.ncbi.nlm.nih.gov/15234467/
**[54]**
No mitochondrial haplotype was found to increase risk for Alzheimer's disease. Biol Psychiatry (1998)
Zsurka, G, J Kalman, A Csaszar, I Rasko, Z Janka, P Venetianer.
https://pubmed.ncbi.nlm.nih.gov/9755361/
**[55]**
Mitochondrial DNA haplogroups and susceptibility to AD and dementia with Lewy bodies. Neurology (2000)

Chinnery, PF, GA Taylor, N Howell, RM Andrews, CM Morris, RW Taylor, IG McKeith, RH Perry, JA Edwardson, DM Turnbull.
https://pubmed.ncbi.nlm.nih.gov/10908912/
**[56]**
Mitochondrial DNA haplogroup cluster UKJT reduces the risk of PD. Ann Neurol (2005)
Pyle, A, T Foltynie, W Tiangyou, et al.
https://pubmed.ncbi.nlm.nih.gov/15786469/
**[57]**
Lack of association between mtDNA haplogroups and Alzheimer's disease in Tuscany. Neurol Sci ( 2007)
Mancuso, M, M Nardini, D Micheli, et al.
https://link.springer.com/article/10.1007/s10072-007-0807-z
**[58]**
Mitochondrial DNA haplogroups in early-onset Alzheimer's disease and frontotemporal lobar degeneration. Mol Neurodegener (2010)
Kruger, J, R Hinttala, K Majamaa, AM Remes.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2830999/
**[59]**
No consistent evidence for association between mtDNA variants and Alzheimer disease. Neurology.( (2012)
Hudson, G, R Sims, D Harold, et al.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317529/

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

# DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe entregarse en la Secretaría de la EINA, dentro del plazo de depósito del TFG/TFM para su evaluación).

D./Dª. Juan Asensio Ayesa ,en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza, Declaro que el presente Trabajo de Fin de (Grado/Máster) Grado

(Título del Trabajo)

Asociación de haplotipos mitocondriales con biomarcadores estructurales de MRI para la caracterización de la enfermedad de Alzheimer (Association of mitochondrial haplotypes with MRI structural biomarkers for the characterization of Alzheimer's disease)

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, a 20 de Noviembre de 2020

Fdo: Juan Asensio Ayesa