universidad de león

TESIS DOCTORAL

Evaluación del rendimiento de metodologías univariantes, multivariantes y de aprendizaje automático en el análisis de variaciones genómicas

Autor:
Fidel DÍEZ DÍAZ

Tutor:
Dr. Vicente MARTÍN SÁNCHEZ
Directores:
Dr. Vicente MARTÍN SÁNCHEZ
Dr. Fernando SÁNCHEZ LASHERAS

Memoria presentada para aspirar al grado de Doctor en BIOMEDICINA Y CIENCIAS DE LA SALUD

Gijón, 3 de abril de 2023

Declaración de autoría

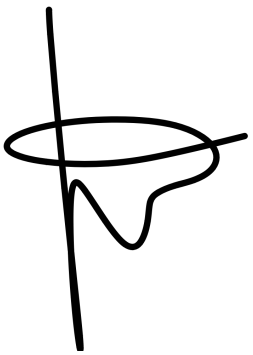
D., Fidel DÍEZ DÍAZ, declara que la memoria de la tesis presentada bajo el título, «Evaluación del rendimiento de metodologías univariantes, multivariantes y de aprendizaje automático en el análisis de variaciones genómicas» es, conforme al artículo 13.1 del R.D. 99/2011, de 28 de enero, un trabajo original de investigación, sin contribución significativa de otra persona que no aparezca reflejada en al misma, y citando adecuadamente la procedencia del contenido no original, conforme a la normativa vigente.

Asimismo, declaro que este trabajo no ha sido presentado y no lo será en el futuro como tesis doctoral, en ninguna universidad o institución de investigación, en España o en el extranjero.

Entiendo la política de tolerancia cero frente al plagio de la Universidad de León, la cual se reserva el derecho a retirar mi título de doctor y adoptar cuantas medidas procedan legalmente, en caso de incumplimiento de este compromiso.

Gijón a 31 de marzo de 2023

Fdo:



DIEZ DIAZ FIDEL - 76956446X 2023.04.03 13:00:11 +02'00'

universidad de león

esDule
Escuela de Doctorado de la Universidad de León

INFORME DEL TUTOR

El Dr. D. Vicente Martín Sánchez como Tutor de la Tesis Doctoral titulada "Evaluación del rendimiento de metodologías univariantes, multivariantes y de aprendizaje automático en el análisis de variaciones genómicas" realizada por D. Fidel Díez Díaz en el programa de doctorado de Biomedicina y Ciencias de la Salud, regulado por el R.D. 99/2011, de 28 de enero, informa favorablemente el depósito de la misma, dado que reúne las condiciones necesarias para su defensa.

Lo que firmo, en León a 31 de marzo de 2023

MARTIN SANCHEZ VICENTE - 10182250N

Firmado digitalmente por MARTIN SANCHEZ VICENTE - 10182250N
Fecha: 2023.04.03 16:43:01 +02'00'

«Sólo podemos ver poco del futuro, pero lo suficiente para darnos cuenta de que hay mucho que hacer.»

Alan Turing

«Lo malo de hacer sugerencias inteligentes es que uno corre el riesgo de que se le asigne para llevarlas a cabo.»

Groucho Marx

«Me gusta pensar en todas esas personas que me enseñaron cosas que nunca había imaginado antes.»

Charles Bukowski

«Y de beber, albóndigas.»

Homer Simpson

Agradecimientos

Me gustaría expresar mi más sincero agradecimiento a mis directores de Tesis, el Dr. Vicente Martín y el Dr. Fernando Sánchez, por su paciencia y dedicación a lo largo de todo este proceso. Sus experiencias y conocimientos me han sido de un valor incalculable y gracias a ellos he podido alcanzar mis objetivos. Ha sido un verdadero honor trabajar a su lado, y estoy muy agradecido por todo lo que han hecho por mí.

Además, quiero dedicar un agradecimiento especial a mi familia, por su apoyo, comprensión y su confianza en mí. Han sido mi fuente de motivación durante todo este camino, y no hubiera podido lograr esto sin su incondicional respaldo.

Por último, pero no menos importante, quiero agradecer a mi hijo Teo, por ser mi luz y mi motor en todo momento. Su sonrisa, su cariño y su inocencia me han dado la fuerza necesaria para superar los obstáculos y seguir adelante en los momentos más difíciles. Espero que este trabajo pueda servir como un ejemplo de perseverancia y dedicación para él.

Gracias a todos ellos, hoy puedo decir que he alcanzado una de las metas más importantes de mi vida.

Índice general

Índice de figuras

Índice de tablas

List of Abbreviations

ADN	Ácido desoxirribonucleico
AG	Algoritmo genético
AMPK	AMP-activated protein kinase
AUC	Área bajo la curva ROC
CCR	Cáncer colorrectal
CORECT	Colorectal cancer transdisciplinary study
DCV	Disease common variant
GASVeM	Genetic algorithms support vector machines
GBM	Gradient boosting machines
GO	Gene ontology
GWAS	Genome wide association studies
HPO	Human phenotype ontology
JSNP	Japanese single nucleotide polymorphism database
KEGG	Kyoto encycclopedia of genes and genomes
LASSO	Least and shrinkage and selection operator
MAF	Frecuencia del menor alelo
NHGRI	Instituto Nacional de Investigación del Genoma Humano
OMIM	Online mendelian inheritance in man
OPEN	Objective prioritization for enhanced novelty
QQ	Cuantil-cuantil
ROC	Receiver operating characteristic curve
SM	Síndrome metabólico

SNP Single nucleotide polymorphism

SVM Máquina de vectores de soporte

WOS Web of Science

Dedicado a mi hijo.

Capítulo 1

Introducción

1.1. Los principios del análisis de la asociación genética

La asociación genética se puede definir como la ocurrencia simultánea de herencias o de características con una probabilidad superior a la que sería esperable por azar [1]. El estudio de la asociación genética trata de identificar esas relaciones con el propósito de establecer un vínculo con un fenotipo observable, como, por ejemplo, una enfermedad, que puede aportar información acerca del mecanismo que lleva a la aparición de cierto rasgo. Además debe tenerse en cuenta que puede existir una asociación entre polimorfismo genéticos (existencia de dos o más variantes) bien por su proximidad física (ligadura genética) [2] o debido al desequilibrio de ligamiento.

La proximidad física entre cromosomas es un factor importante en lo referente a los principios claves de la asociación genética. Dado que los puntos de recombinación genética son esencialmente aleatorios, una mayor distancia entre alelos incrementa la probabilidad de que estén separados y viceversa, dado que existen ciertos alelos que dan lugar a fenotipos fácilmente observables.

La consecución del primer mapa completo del genoma humano, secuenciado gracias al proyecto International Human Genome Project (1990-2003) [3] supuso un hito en la investigación en ciencias de la salud. Más concretamente, originó avances en el análisis de la ligadura genética, donde los marcadores de genotipos en los genomas de los pacientes sirvieron para la identificación de ciertas enfermedades monogénicas [1]. Sin embargo, también gracias a este tipo de estudios se puso de manifiesto que los genes relacionados con multitud de enfermedades no se podían determinar haciendo uso de este método. El motivo de esto es que muchas de las enfermedades comunes son multifactoriales. Este descubrimiento condujo a la formulación de la hipótesis conocida como *Disease-Commom Variant* (CDCV) [4]. Esta hipótesis propone que las causas de las enfermedades comunes son algunos polimorfismos presentes con alta frecuencia dentro de la población y que se originaron como mutaciones en antepasados comunes, habiendo sido heradadas por sus descendientes.

1.2. La estratificación de la población y sus consecuencias

La estratificación de la población se define como la existencia de diferencias sistemáticas en las frecuencias alélicas que surgen debido a diferencias en la ascendencia de las subpoblaciones consideradas en un estudio dado [5]. Así, lógicamente, las subpoblaciones con bajas frecuencias de apareamiento entre ellas están sujetas a diferencias en su deriva genética, de manera que las frecuencias de los alelos que no

están bajo presión selectiva pueden divergir por casualidad, dado el tiempo suficiente. Por tanto, estas diferencias de ascendencia pueden confundir los verdaderos determinantes genéticos subyacentes al fenotipo de interés [6]. Por este motivo, es importante controlar la estratificación de la población con el fin de identificar las verdaderas asociaciones. Una forma obvia de controlar la estratificación sería asegurar una completa homogeneidad de la población durante la etapa de diseño experimental, por ejemplo, mediante el uso de información de origen étnico o ascendencia familiar en la etapa de reclutamiento para el estudio. Aunque ésta sigue siendo una de las formas más importantes de controlar la estratificación, suele estar sujeta a inexactitudes considerables y, a menudo, resulta insuficiente para reflejar plenamente la complejidad total de la posible estructura de la población. Alternativamente, se puede utilizar un diseño basado en la familia, donde los datos se recopilan de individuos de los que se sabe a ciencia cierta que están relacionados y, por lo tanto, se garantiza que no se verán afectados por problemas de estratificación de la población.

La detección y cuantificación de la estratificación de la población es posible utilizando el método de control genómico propuesto por Devlin y Roeder [7]. Este método utiliza una prueba de tendencia de Cochran-Armitage [8, 9] para calcular el factor de inflación, que luego puede usarse para ajustar las estadísticas de prueba de asociación relevantes. Sin embargo, una desventaja es que no se tienen en cuenta las posibles diferencias entre alelos individuales, ya que el ajuste se aplica de manera uniforme. Para permitir una mayor flexibilidad, se propusieron pruebas de asociación estructuradas [10] que buscan identificar subgrupos o grupos de individuos y, por lo tanto, permiten una mayor flexibilidad, pero su aplicación es costosa desde el punto de vista computacional y dependen de parámetros adicionales, como el número de clústeres.

Con el fin de superar estas limitaciones, se desarrolló otro método alternativo, que utiliza el análisis de componentes principales para capturar la estructura de la población [11]. El análisis de componentes principales identifica los principales ejes de variación dentro de los datos y ha demostrado ser capaz de reflejar con precisión la etnia o incluso la distancia geográfica entre las muestras. La cantidad de variación atribuida a ciertos ejes particulares se puede utilizar directamente para ajustar los efectos de la estratificación de la población, incorporándolos como covariables en un modelo de regresión utilizado para la prueba de asociación a nivel de muestras individuales. Debido a su gran eficiencia y flexibilidad computacional, el análisis de estratificación basado en componentes principales es, en la actualidad, el método más comúnmente utilizado para controlar la estratificación de la población [1]. Por lo general, se utiliza un subconjunto de marcadores de referencia para realizar el análisis e identificar cualquier muestra atípica altamente divergente, que luego se excluye. Si aún se determina que el conjunto de datos restantes está sujeto a una estratificación sustancial, los componentes principales se agregan al modelo como una forma simple y eficiente de ajustar esos efectos.

1.3. Los estudios de genoma amplio

Los resultados del Proyecto Genoma Humano [12] y el Proyecto Internacional HapMap [13] hicieron posible, hace ya dos décadas, encontrar genes relacionados con rasgos y enfermedades. Así, los estudios de asociación de genoma amplio que se han realizado desde entonces, han conseguido descubrimientos de gran interés

en genética humana.

Los estudios de genoma amplio, conocidos en inglés como *genome wide association studies* (GWAS), tienen como objetivo el estudio de las variaciones en el ADN, fundamentalmente las variaciones en los mononucleótidos de polimorfismo único, conocidos como SNPs (*single nucleotide polymorphisms*) del genoma [1] con el fin de conocer cómo éstas afectan al riesgo del padecimiento de ciertas enfermedades, la respuesta a ciertos tratamientos de cada paciente, la presencia o no de cierto rasgo en un individuo, etc. En la actualidad, los estudios de genoma amplio se han convertido en una de las herramientas más potentes de las que se dispone para entender la genética humana. El primer estudio que probó la utilidad de los GWAS fue el publicado en 2002 por Ozaki y cols. [14] que identificaron algunas de las variantes genéticas asociadas con el infarto de miocardio.

Casi a la vez, en el año 2001, surgió el proyecto de colaboración internacional HapMap [13], con el que se desarrolló una base de datos de información genética con el objetivo de poder realizar estudios GWAS. Este proyecto se basó en dos hallazgos científicos ya conocidos. Por una parte, las variantes genéticas son distintas de una etnia a otra y, por tanto, la frecuencia de alelos de algunos *loci* puede ser distinta, lo que permite la estratificación de la población y la mejora de la atención médica. Por tanto, resulta necesario disponer de información de sujetos de las distintas etnias implicadas en el estudio para el grupo de control. El segundo, es el conocido como desequilibrio de ligamiento [15]. Se trata de la asociación no aleatoria de alelos en diferentes *loci* dentro de una población donde se considera que los *loci* se encuentran en desequilibrio de ligamiento cuando la frecuencia de asociación de sus diferentes alelos es mayor o menor de lo esperado si los *loci* fueran independientes y se asociaran de forma aleatoria. Suele ocurrir en genes que se encuentran en el mismo cromosoma y próximos entre sí. La existencia del desequilibrio de ligamiento motiva que en algunas ocasiones, algunos SNPs se encuentren completamente ligados a otros. En aquellos casos en los que dos SNPs se encuentran completamente ligados, esto se debe de tener en cuenta a la hora de llevar a cabo los estudios GWAS.

En la actualidad, una de las principales críticas que se hace a los estudios GWAS es que hasta la fecha, la mayoría de los descubrimientos no se han aplicado a la práctica clínica [16], pero a pesar de este evidente inconveniente, los GWAS sí tienen una gran relevancia. A modo de ejemplo, se puede decir que hasta el desarrollo de los estudios GWAS no había sido posible encontrar ningún gen ligado a la esquizofrenia [17].

Los GWAS no solo son de interés para el descubrimiento de asociaciones sólidas, sino que también brindan información sobre la naturaleza de las variaciones en los rasgos y han contribuido al descubrimiento de nuevos conocimientos biológicos sobre cómo las variaciones del ADN pueden afectar la regulación genética. Las variaciones en el genoma humano se deben principalmente a dos causas: mutaciones puntuales y cambios de la estructura [18]. Cuando ocurre una mutación puntual, una base de ADN es reemplazada por otra. En este caso de variaciones estructurales como esta, los cambios son más amplios y pueden variar desde pequeñas inserciones o deleciones hasta grandes reordenamientos cromosómicos [18]. Cada tipo de variación estructural tiene diferentes tasas de mutación y evolución y su papel en la variación fenotípica es en la mayoría de los casos desconocida.

1.4. Los primeros estudio GWAS

Con el fin de identificar los genes que se relacionan con las enfermedades comunes a través de los estudios GWAS, en primer lugar, fue necesario aislar los polimorfismos de un solo nucleótido, conocidos por sus siglas en inglés SNPs (*single nucleotyde polimorphism*). Los SNPs representan las variantes genéticas más comúnmente encontradas en el genoma humano. Dada su amplia distribución, estos polimorfismos se localizan en cualquier parte de la estructura de los genes y el genoma [19]. Uno de los primeros proyectos que se llevaron a cabo con el fin de recoger esta información, fue el acometido por el Institute of Medical Science de Japón y la Universidad de Tokio con el apoyo de la Agencia Japonesa de Ciencia y Tecnología (2000-2002) [20].

Durante la realización del trabajo mencionado, se secuenció el ADN de 24 individuos. El análisis de éste identificó un total de 174 269 polimorfismos que fueron puestos a disposición de la comunidad científica a través de la denominada *Japonese Single Nucleotide Polymorphisms Database* (JSNP) (http://snp.ims.u.tokio.ac.jp). Dentro del marco de este proyecto, se desarrolló un sistema robotizado que permitía realizar el genotipado de una forma más precisa de lo que se había conseguido hasta ese momento (método INVADER) [20], que fue de gran utilidad para la recolección de información.

Desde el punto de vista biomédico, uno de los hallazgo más importante fue, el realizado en 2002 consistente en el descubrimiento de genes relacionados con el infarto de miocardio [14]. Además, el análisis a gran escala de unos 80000 SNPs en 564 individuos condujo al desarrollo del primer mapa LD/haplotype blocks de todos los cromosomas humanos, lo que contribuyó a mejorar la eficiencia de los subsiguientes esfuerzos de genotipado, identificando conjuntos representativos de SNPs que capturaban información suficiente de los haplotipos de unos 13000 genes [6]. La puesta a punto de estas metodologías permitieron un rápido avance en el descubrimiento de genes relacionados con múltiples enfermedades.

Otros dos de los primeros estudios GWAS que se realizaron y resultaron ser de gran interés, fueron los publicados en 2005 y 2006 [21, 22]. Ambos trabajos encontraron variantes comunes asociadas a la degeneración macular relacionada con la edad. Nótese que un estudios GWAS puede ir más allá de los estudios de genes candidatos. Las razón de porqué la capacidad analítica de los estudios GWAS es mayor, es debido fundamentalmente a dos motivos. Por un lado, mientras que en los estudios de genes candidatos solo se consideran unos pocos polimorfismos de nucleótido único, un estudio GWAS significa estudiar simultáneamente un gran número de SNPs representativos de la variación genética del genoma completo; También vale la pena señalar que los GWAS se consideran "libres de hipótesis". En otras palabras, son capaces de buscar efectos de riesgo comunes al observar los SNPs ubicados en toda o, al menos, en una parte considerable del genoma sin ninguna lista de loci a priori [23].

1.5. El diseño experimental de los estudios GWAS

Los estudios GWAS son capaces de analizar el papel de las variaciones genéticas comunes en enfermedades humanas complejas. En un principio, se esperaba que los

estudios GWAS tuvieran la ventaja de no depender del conocimiento previo de las vías biológicas en comparación con los estudios de *genes candidatos* [24]. Esta ventaja permite a los estudios GWAS superar el sesgo de los estudios de *genes candidatos*. Nótese que los estudios de gen candidato se centran en las asociaciones entre la variación genética dentro de genes de interés preespecificados y los fenotipos. Las vías biológicas se pueden definir como un grupo de genes que están relacionados desde un punto de vista funcional.

El principal desafío del análisis de datos que se desarrolla en los estudios GWAS es la arquitectura poligénica de enfermedades complejas. Esto significa que en presencia de muchas variantes con efectos pequeños o moderados, se necesita un tamaño de muestra grande tanto para el mapeo de asociación como para la predicción del riesgo. Sin embargo, la selección de muestras puede resultar costosa y requerir mucho tiempo. El paso clave para la validación de la asociación entre variantes genéticas y enfermedades humanas complejas es la replicación de los hallazgos en muestras independientes. Por lo general, se considera que la replicación de asociaciones informadas es la mejor forma de validar los resultados obtenidos en los estudios GWAS.

En los primeros estudios GWAS que se realizaron, se consideraba la existencia de un vínculo hipotético entre los SNPs y el fenotipo, pero estudiando únicamente la relación de un SNP de cada vez [25]. En la actualidad, los estudios hacen uso de análisis más complejos que incluyen análisis multivariante [26] o enfoques de aprendizaje automático [27, 28]. Gracias al uso de estudios GWAS se ha conseguido una mejor comprensión de las componentes genéticas de muchos rasgos complejos.

Con el paso de los años, las metodologías de los estudios GWAS para asociar estas variantes genéticas con enfermedades y otros rasgos fenotípicos, se han vuelto cada vez más refinadas y se han estandarizado. Como ocurre con otras metodologías biomédicas con base estadística, resulta imprescindible contar con un diseño experimental apropiado para asegurar el éxito de estos estudios. El motivo es que los riesgos relativos hipotéticos atribuibles a factores particulares, el número de muestras y el número de marcadores a examinar (en relación con pruebas múltiples) afectan directamente el poder estadístico, es decir, el problema de los análisis GWAS es trabajar con errores alfa muy pequeños, con lo que se incrementa el error beta. La estimación del poder estadístico y la determinación del tamaño de la muestra necesario para detectar un efecto de asociación significativo se realizan comúnmente probando la diferencia en las proporciones de población relevantes [29].

Con frecuencia se realiza una prueba estadística independiente para verificar una asociación entre cada *locus* genético individual y un fenotipo de interés. Por lo tanto, el número de tales pruebas puede ascender fácilmente a millones. Nótese que la contabilidad adecuada de las pruebas múltiples es particularmente importante para controlar los resultados de falsos positivos. Desde la perspectiva del diseño experimental, conviene destacar el papel central que desempeña la replicación de resultados para garantizar la solidez de todos los hallazgos de un estudio GWAS. Por estas razones, en caso de ser posible, resulta conveniente disponer en los estudios GWAS de al menos dos conjuntos de muestras, un subconjunto de *descubrimiento* y un subconjunto de *replicación*, posiblemente más pequeño. Idealmente, el subconjunto de *replicación* se debería generar a partir de una cohorte distinta de pacientes. Luego, estos conjuntos de muestras se analizan de forma independiente y se comparan los

resultados significativos de todo el genoma, con la idea de que solo las variantes que se detectaron en ambos conjuntos de datos representan asociaciones verdaderas.

1.6. La verificación estadística de la asociación genética

Una vez que se ha preparado la información genética para el estudio y se han realizado los controles de calidad necesarios, el siguiente paso consiste en investigar aquellas asociaciones genéticas que puedan explicar el fenotipo de interés observado. Normalmente, un rasgo vinculado a un *locus* particular puede ser binario, como el padecer cierta enfermedad o no, para lo que resulta muy adecuado un estudio de casos y controles, o bien tratarse de un rasgo cuantitativo, como la altura o los niveles de colesterol. Una asociación de particular interés entre las variables cuantitativas es la que vincula la variación genética con los patrones de expresión de genes particulares, denominada *Quantitative Trait Locus* (eQTL). Según el diseño de un estudio en particular, los individuos reclutados pueden pertenecer a una serie de familias conocidas o bien considerarse como no relacionados. En aras de la brevedad, esta sección solo se ocupará del diseño de estudio más común en el que los individuos reclutados no están relacionados, lo que se denomina un diseño de estudio "basado en la población", como es el caso del que ha sido empleado en el presente proyecto de investigación.

Esta sección describe las estrategias más típicas para identificar asociaciones de variantes comunes. En el escenario más típico, los alelos particulares no conducen necesariamente a cierta manifestación de un rasgo binario, sino que alteran la probabilidad o el riesgo de que ocurra. La probabilidad de que un individuo de una población muestre un rasgo se denomina formalmente "penetrancia". Dado que en un genoma humano diploide normalmente están presentes dos posibles copias de cada alelo, el modelo estadístico correcto de la relación entre genotipo y fenotipo dependerá del tipo de dominancia genética en un locus dado. Asimismo, el número de alelos puede tener un efecto aditivo o multiplicativo, y esto es igualmente aplicable tanto para la magnitud de los rasgos cuantitativos como para la penetrancia de los rasgos binarios. En el caso de los rasgos binarios, la fuerza de la asociación se puede cuantificar como una razón de probabilidades"para un rasgo de interés dados los genotipos alternativos particulares.

Teniendo en cuenta la naturaleza computacionalmente intensiva del análisis a una escala de genoma completo, cada locus generalmente se prueba de forma independiente de todos los demás. En un escenario más simple, se llamarán genotipos exactos y, por lo tanto, en el caso de una prueba binaria, los datos se pueden representar como una tabla de contingencia donde los recuentos en cada celda serían números de individuos con una categoría de combinación de genotipo-rasgo particular. El tipo de modelo determina cómo se construye la tabla. Así, por ejemplo, se empleará una tabla de dos por dos en el caso de un modelo dominante o recesivo o una de dos por tres si no se asume ningún modelo en particular. Como generalmente no se conoce el modelo correcto, es común asumir un modelo aditivo, que se puede representar con una tabla de contingencia de dos por tres que también se considera que tiene una relación ordenada con el rasgo. Si existe una suposición de tendencia u orden, esta relación se puede capturar usando la prueba de tendencia de Cochran-Armitage; de lo contrario, se puede usar una prueba de Chi-cuadrado [30] si se considera más apropiada la independencia entre todas las categorías. Sin

embargo, en la práctica, a menudo es muy conveniente incorporar covariables en el modelo que estos tipos de pruebas simples no pueden tener en cuenta. Por ejemplo, la probabilidad de desarrollar enfermedades particulares a menudo aumenta con la edad o puede verse afectada por el sexo de la persona. Esta información solo puede incorporarse mediante el uso de modelos multivariados más sofisticados. En el caso del análisis GWAS, los modelos de regresión logística [31] o lineal [32] son los más utilizados. Por lo general, los modelos de regresión logística para rasgos binarios incluyen covariables para la edad, el género y, cuando se corrige la estratificación de la población, los primeros valores de los componentes principales de cada muestra.

Dada la complejidad de los estudios GWAS y la gran cantidad de factores que potencialmente pueden conducir a sesgos, es de vital importancia verificar e identificar la posible presencia de alguno de estos problemas en los estudios. Una forma genérica que se usa comúnmente para verificar los resultados son los gráficos de cuantil-cuantil (QQ) [33] de los valores finales de significancia de la asociación. Dado que, generalmente, el número de hallazgos de interés en un GWAS es pequeño, se espera que los patrones de los SNPs no relacionados sean efectivamente aleatorios, es decir, la probabilidad de observar un valor de significancia particularmente alto por casualidad solo está influenciada por el número de muestras en un conjunto de datos. Así, un gráfico QQ es un gráfico de dispersión de los valores de significación esperados frente a los observados, que se pueden utilizar para verificar este patrón. Si se han tenido en cuenta todas las fuentes de sesgo, la mayoría de los puntos caerían en una línea de 45 grados, con un pequeño conjunto de puntos muy significativos por encima de esta línea si está presente una verdadera señal de asociación. Este análisis a menudo también se resume como una estadística del factor de inflación genómica [34]. El factor de inflación genómica se define formalmente como una relación entre las medianas de distribución Chi-cuadrado real sobre la esperada. Cuando el valor es cercano a uno, esto significa que no hay inflación. Aunque lo más habitual es que este análisis se utilice para comprobar la presencia de una subestructura de población, también se pueden detectar otros tipo de efectos, como los efectos de bloque.

Teniendo en cuenta que el número de *loci* que pueden perfilarse mediante la secuenciación del genoma completo o tecnologías equivalentes respaldadas por la imputación del genoma puede ser de millones, es particularmente importante corregir los valores de significación para el número de pruebas realizadas. Sin embargo, los procedimientos estándar para corregir la tasa de error familiar, como la corrección de Bonferroni [35], asumen independencia entre las pruebas individuales. Debido a los patrones de desequilibrio de ligamiento, esta suposición no es cierta en el caso de los estudios de genoma amplio, conocidos como GWAS (genome-wide association studies) y, por lo tanto, es probable que dichos métodos sean demasiado conservadores [36]. Estimaciones previas determinaron que un número apropiado de supuestas señales independientes está aproximadamente en la región de 1,000,000. Los conocimientos de este trabajo se utilizaron para derivar un límite de significación GWAS ampliamente aceptado de $5 \cdot 10^{-8}$, aunque debe tenerse en cuenta que esta estimación es más aplicable para la población europea y el valor correcto verdadero dependería de la diversidad de la población objeto de estudio. Otra alternativa es utilizar una prueba de permutación para calcular los valores de significación ajustados. Al permutar las etiquetas de respuesta y calcular la importancia, se puede calcular una distribución empírica de probabilidades. Por esta razón, la prueba de permutación se considera el mejor método de corrección, sin embargo, este enfoque es muy intensivo desde el punto de vista computacional, lo que puede hacer que no

sea factible aplicarlo en la práctica, aunque la eficiencia puede mejorarse utilizando métodos aproximados [37].

Para garantizar la veracidad de las asociaciones encontradas, el último paso del análisis suele consistir en la replicación del resultado en un conjunto de datos independiente. La replicación es particularmente importante en el contexto de los estudios GWAS, ya que se ha encontrado que los patrones genómicos subyacentes a los fenotipos poligénicos tienden a ser muy complejos y es común identificar un gran número de *loci* que explican individualmente solo una cantidad muy pequeña de la heredabilidad total. El tamaño del efecto observado puede significar que los estudios a menudo no tienen el poder estadístico suficiente para confirmar de manera satisfactoria el verdadero efecto. Asimismo, la replicación puede ayudar a identificar y descartar asociaciones espurias que surgen debido al sesgo y también puede servir para confirmar la existencia del efecto en diferentes conjuntos de condiciones y derivar una estimación más precisa del tamaño real del efecto.

1.7. Avances metodológicos recientes en el análisis de la asociación de genotipos

La heredabilidad genética se refiere a la velocidad a la que un fenotipo particular es heredado por la descendencia de cierto progenitor. Al comparar la heredabilidad conocida, por ejemplo, con qué frecuencia los hermanos heredan una enfermedad de sus padres, es posible determinar qué parte de la variación en un fenotipo se explica por los polimorfismos genéticos identificados actualmente. El análisis de GWAS convencional considera los efectos individuales de los polimorfismos genéticos sobre el rasgo de interés. Sin embargo, en la actualidad se ha hecho evidente que la totalidad de tales variaciones todavía explica solo una pequeña parte de toda la heredabilidad conocida. Este fenómeno se conoce como el problema de la *falta de heredabilidad* [38]. Se han propuesto varias explicaciones para este problema, incluidas las posibles limitaciones metodológicas para estimar la heredabilidad verdadera, así como para medir o definir con precisión los fenotipos y los posibles efectos epigenéticos. Otras posibles explicaciones atribuyen la heredabilidad faltante a efectos genómicos que no son capturados adecuadamente por los métodos de análisis clásicos de GWAS, como interacciones, influencia de polimorfismos raros o efectos altamente poligénicos [1]. Si un fenotipo está determinado por los efectos aditivos de un gran número de polimorfismos con efectos individuales muy pequeños, el simple hecho de aumentar el número de muestras aumentará suficientemente el poder estadístico para detectar todas estas pequeñas asociaciones, aunque esta estrategia, inevitablemente, estará sujeta a rendimientos decrecientes. Por el contrario, otras posibilidades implican que el problema de la heredabilidad faltante puede eventualmente resolverse mediante mejoras adicionales en la metodología y ya se han propuesto varios enfoques novedosos para explorar estas vías.

Se han realizado avances considerables en la detección de los efectos de los SNPs raros. La detección de polimorfismos raros mediante estudios GWAS convencionales, conduce a un riesgo inflado debido a las detecciones de falsos positivos, como consecuencia de frecuencias de alelos altamente desequilibradas que violan los supuestos de distribución de las pruebas de significación comúnmente utilizadas. En la mayoría de los procesos de análisis actuales, este riesgo se mitiga al no considerar ningún polimorfismo donde la frecuencia de los alelos menores esté por debajo

de un umbral particular. Los valores más comúnmente empleados son el 1 % o 5 % de todas las muestras perfiladas en el estudio. Para capturar estos efectos, los SNP raros se pueden combinar y considerar como un grupo, donde una prueba consideraría el efecto general de un conjunto de polimorfismos [39], generalmente en el contexto de alguna forma de modelo. En consecuencia, tales pruebas requieren información adicional sobre cómo agrupar diferentes SNPs en conjuntos significativos. Una estrategia habitual consiste en agrupar los SNPs en torno a genes o incluso vías particulares.

La interacción entre polimorfismos ocurre cuando el efecto de un alelo depende condicionalmente del efecto de otro, un fenómeno también conocido como epistasis [40, 41]. La detección de interacciones es un desafío debido a su naturaleza combinatoria, lo que significa que se requiere de un gran número de pruebas individuales para verificar exhaustivamente todas las posibilidades [42].

Los métodos de detección de epistasis comúnmente implican el desarrollo de estrategias para reducir el número de pruebas realizadas mediante el uso de alguna forma de conocimiento previo, por ejemplo, al observar las interacciones entre polimorfismos que se consideran individualmente significativas. Finalmente, cabe señalar que es muy probable que alguna combinación de estas posibles explicaciones subyazca al problema de la falta de heredabilidad y se ha encontrado alguna evidencia que sugiere la influencia de todos estos factores en casos particulares. También es probable que diferentes factores sean prominentes para diferentes tipos de fenotipos. Dada esta diversidad de hipótesis posibles y la ausencia de una solución definitiva, en la actualidad la pregunta sobre las causas de la falta de heredabilidad y las mejores estrategias para abordarla sigue siendo objeto de activo debate.

1.8. El aprendizaje automático

El aprendizaje automático o *machine learning* es el subcampo de las ciencias de la computación y rama de la inteligencia artificial cuyo objetivo es el desarrollo de técnicas que permitan el aprendizaje de las máquinas, entendiendo como tales, fundamentalmente, las basadas en dispositivos electrónicos. Es decir, lo que se busca en este campo del conocimiento es el desarrollo de metodologías que permitan, a partir de conjuntos de datos de referencia, la generalización de comportamientos y, por tanto, otorguen a la máquina un comportamiento en apariencia inteligente. Según Russell y Norvig [43], el aprendizaje automático puede interpretarse como un intento de automatización parcial del método científico a través de las matemáticas. Por lo tanto, se habla de un proceso de inducción del conocimiento. Los orígenes del aprendizaje automático se remontan a la década de 1940; más concretamente a 1943, año en el que McCulloch y Pitts publicaron el primer modelo matemático relativo al funcionamiento de las neuronas [44]. Pocos años después, Turing [45] enunció la desde entonces conocida como prueba de Turing. Se considera que una máquina es capaz de pasar la prueba de Turing si un humano, hablando con ella y con otro humano simultáneamente, no es capaz de distinguir cuál de sus dos interlocutores es la máquina. En esa misma década Samuel (1959) [46] desarrolló un programa que era capaz de jugar a las damas basándose, para la elección de cada movimiento, en las consecuencias que dicha elección tendría en las jugadas futuras. La primera red neuronal artificial funcional, denominada perceptrón, fue diseñada por Rosenblatt

en 1958 [47] con el fin de reconocer formas y patrones. Otro de los primeros experimentos relacionados con las redes neuronales artificiales fue el que desarrollaron Widrow y Hoff, de la Universidad de Stanford, quieres crearon en 1959 ADALINE, una red capaz de detectar patrones binarios y de predecir cuál sería el valor del siguiente bit [48]. La siguiente red neuronal que desarrollaron fue, en 1962, la denominada MADELINE que era capaz de eliminar el eco en conversaciones telefónicas [48]. Desde entonces, el progreso de las diferentes metodologías de machine learning ha continuado, y su implementación ha estado siempre condicionada por el avance de las capacidades de cálculo de los ordenadores.

1.8.1. Las matemáticas del aprendizaje automático

Galileo Galilei (1623) en su obra «El ensayador» [49] afirmó:

> «La filosofía está escrita en ese grandísimo libro que tenemos abierto ante los ojos, quiero decir, el universo, pero no se puede entender si antes no se aprende a entender la lengua, a conocer los caracteres en los que está escrito. Está escrito en lengua matemática y sus caracteres son triángulos, círculos y otras figuras geométricas, sin las cuales es imposible entender ni una palabra; sin ellos es como girar vanamente en un oscuro laberinto».

En el caso del aprendizaje automático, el uso de las matemáticas es una necesidad inherente al propio conocimiento de esta disciplina. Sin entrar en métodos concretos, desde un punto de vista conceptual, en *machine learning*, según Russell y Norvig [43], existen tres preguntas fundamentales:

- ¿Cuáles son las reglas formales que conducirán a la obtención de conclusiones válidas?
- ¿Qué es susceptible de ser computado?
- ¿Cómo se deben de llevar a cabo los razonamientos cuando la información disponible es imprecisa?

La formalización de todos los conceptos relacionados con el aprendizaje automático requiere del empleo de las matemáticas, haciendo uso fundamentalmente de cuatro áreas: la lógica, la computación, la probabilidad y el análisis numérico. En lo relativo a la lógica, la disciplina del aprendizaje automático es deudora de los trabajos de Boole que, en su obra titulada «The Mathematical Analysis of Logic, Being an Essay towards a Calculus of Deductive Reasoning» [50], pone las bases de la lógica proposicional o booleana, así como de la extensión de la misma por Frege, el cual incluyó objetos y relaciones creando la lógica de primer orden que se sigue empleando en la actualidad [51], al igual que de los trabajos de Tarski [52] que contribuyeron a la madurez de la lógica de primer orden.

El primer teorema de incompletitud de Gödel (1931) [53] afirma que:

> «Bajo ciertas condiciones, ninguna teoría matemática formal capaz de describir los números naturales y la aritmética con suficiente expresividad es, a la vez, consistente y completa».

Es decir, si los axiomas de dicha teoría no se contradicen entre sí, entonces existen enunciados que no se pueden probar ni refutar a partir de ellos. En particular, la conclusión del teorema se aplica siempre que la teoría aritmética en cuestión sea

recursiva, esto es, una teoría en la que el proceso de deducción se puede llevar a cabo mediante un algoritmo.

El segundo teorema de incompletitud es un caso particular del primero y afirma que una de las sentencias indecibles de dicha teoría es aquella que afirma la consistencia de esta. Es decir, que la consistencia del conjunto de axiomas no se puede deducir con el mero uso de dicho conjunto.

Los teoremas de Gödel se pueden interpretar como que existen ciertas funciones ejecutables sobre los números enteros que no pueden ser ejecutadas a través de un algoritmo. Fueron los teoremas de Gödel los que llevaron a Turing a interesarse por caracterizar exactamente qué funciones son computables, aunque este concepto resulta problemático, dado que no se dispone de una definición general de lo que es computación o de lo que se puede considerar como un procedimiento efectivo. Sin embargo, la tesis de Church-Turing [54] afirma que la máquina de Turing es capaz de computar cualquier función computable. A pesar de no tratarse de un teorema, pues se trata de una afirmación formalmente indemostrable, tiene una aceptación universal. Turing también demostró [55] que existen algunas funciones que no pueden ser computadas por una máquina de Turing.

Otro concepto matemático de especial interés dentro del campo del aprendizaje automático es el de tratabilidad. Dentro de la teoría de la complejidad computacional, se distingue entre los algoritmos de tiempo polinómico y los de tiempo exponencial [56]. Se considera que un algoritmo es computable en tiempo polinomial cuando su función de complejidad temporal se puede representar por $O(p(n))$, siendo p una función polinómica y n el tamaño de la entrada. Todos aquellos algoritmos cuya función de complejidad temporal no puede acotarse de esta forma se denominan algoritmos de tiempo exponencial. En teoría de la complejidad computacional, se considera que un problema no está bien resuelto hasta que se encuentra un algoritmo capaz de resolverlo en un tiempo polinomial. Todo problema que no disponga de un algoritmo de tiempo polinomial para su resolución se considera un intratable.

La teoría de completitud NP proporciona un método que permite reconocer los problemas intratables [56]. Así, la clase de complejidad NP está formada por los problemas verificables en tiempo polinómico. Por verificable se entiende un problema tal que, dado un candidato a solución, se puede verificar que dicho resultado es correcto en un tiempo polinómico en el tamaño de la entrada. A los problemas en la clase NP, usualmente se les denomina problemas NP. El término NP proviene de no determinista en tiempo polinómico.

Otra de las ramas de las matemáticas de gran utilidad en el campo del aprendizaje automático es la teoría de la probabilidad. Cardano, en el siglo XVI, fue el primero en el uso de conceptos probabilísticos para la descripción de los posibles resultados de los juegos de azar [57]. En 1654, Pascal, en una carta a Fermat, afirmaba ser capaz de predecir el resultado de un juego de azar y estimar las ganancias que obtendría cada jugador como resultado de ese pronóstico [58]. El avance de la ciencia estadística se produjo desde ese momento con las aportaciones de otros autores como Bernouilli o Laplace. Entre las contribuciones de mayor interés y más empleadas en la actualidad, se encuentra la de Bayes, quien, con su teorema [59], fue capaz de expresar la probabilidad condicional de un evento aleatorio A dado otro evento

B en términos de la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de solo A.

Finalmente, también cabe señalar que el desarrollo del *machine learning* no sería posible sin el uso de una serie de metodologías propias del análisis numérico. Entre las metodologías del análisis numérico que son de utilidad dentro del campo del aprendizaje automático cabe mencionar el álgebra lineal numérica, la interpolación, la optimización o la transformada rápida de Fourier. Los métodos numéricos se constituyen, por tanto, en herramientas de gran importancia en la resolución de problemas de aprendizaje automático [60].

1.8.2. Clasificación de las metodologías y técnicas fundamentales del aprendizaje automático

Las metodologías propias del aprendizaje automático se pueden dividir en tres categorías principales [60]:

Aprendizaje supervisado. En el aprendizaje supervisado se dispone de la información correspondiente tanto a las variables de entrada como a las de salida. Se entrena el algoritmo con la información de entrada disponible con el fin de que ante un nuevo vector de datos sea capaz de clasificarlo correctamente o bien de asignarle un valor numérico correcto en el caso de que se trate de un modelo de regresión.

Aprendizaje no supervisado. Dentro de la categoría del aprendizaje no supervisado se incluyen todos aquellos métodos en los que los datos no están etiquetados o, dicho de otra forma, en los que únicamente se dispone de información relativa a las variables de entrada, pero no de salida. Este tipo de técnicas son las encargadas de buscar por sí mismas patrones dentro de la estructura de los datos que analizan. El aprendizaje no supervisado persigue el desarrollo de algoritmos capaces de clasificar la información sin intervención externa.

Aprendizaje por refuerzo. Los métodos que se encuadran dentro de la categoría de aprendizaje por refuerzo son aquellos que aprenden a realizar una tarea de forma correcta a través de un sistema de recompensas y castigos. Recompensa cuando ejecutan de forma correcta la tarea encomendada, y castigo cuando lo hacen de forma incorrecta.

Estas tres categorías, si bien son las principales, no son las únicas existentes. Así, por ejemplo, se habla también de aprendizaje semisupervisado, entendiendo como tal aquel que combina el aprendizaje supervisado con el no supervisado para poder hacer una correcta clasificación de la información disponible.

Una lista exhaustiva de todos los métodos de *machine learning* resulta difícil de realizar y está fuera del alcance de este proyecto de investigación. Sin embargo, a continuación, se describen y clasifican los principales métodos existentes:

Agrupamiento o *clustering*. El método de los k vecinos más cercanos. Se trata de una técnica de aprendizaje no supervisada. Es un procedimiento de agrupación de una serie de vectores según criterios, habitualmente, de distancia. En este método se disponen los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Un ejemplo de este tipo de técnicas es el algoritmo

k-medias. El término *k -medias*, en inglés *k-means*, fue empleado por primera vez por MacQueen en 1967 [61]. El algoritmo que se emplea hoy en día para este método fue publicado por Lloyd en 1982 [62]. En esencia, se trata de un método de cuantificación vectorial, originalmente empleado para el procesamiento de señales, que tiene como objetivo dividir n observaciones en k clústeres, en los que cada observación pertenece al clúster con la media más cercana.

Algoritmos evolutivos. Los algoritmos evolutivos son métodos de búsqueda inspirados en los principios de la selección natural y la genética que se pueden usar tanto en un contexto de aprendizaje supervisado como no supervisado. Estos algoritmos evalúan en cada iteración una población de soluciones candidatas, conocidas como individuos, y se quedan con los que proporcionan los mejores resultados, derivando, a partir de estas soluciones, una nueva población. Existen varios tipos de algoritmos evolutivos. Los más conocidos son los denominados genéticos, que se basan en los trabajos de Holland [63] y Goldberg [64].

Árboles de decisión. Se trata de un método de aprendizaje supervisado [60]. Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo [65]. Dada una base de datos, se construye un árbol de relaciones lógicas muy similar a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones sucesivas para la resolución de un problema [66].

Clasificador bayesiano ingenuo. Se trata de una técnica de clasificación supervisada que hace uso del teorema de Bayes para calcular la probabilidad de que un nuevo vector pertenezca a cada una de las clases de datos existentes en el conjunto [65]. El uso del adjetivo ingenuo es debido a que se supone la independencia de las variables.

Redes neuronales. Es un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Como su propio nombre indica, una red neuronal es un sistema de interconexión de neuronas en red de forma que estas colaboran entre sí para producir un estímulo de salida. Algunos ejemplos de redes neuronales son: el perceptrón [67], el perceptrón multicapa [65] y los mapas autoorganizados [68]. Las redes neuronales se pueden emplear dentro de un paradigma de aprendizaje supervisado o no supervisado.

Técnicas de regresión. Los métodos de regresión son técnicas de aprendizaje supervisado que tratan de explicar una variable numérica dependiente a partir de cierto conjunto de variables independientes. Las tres técnicas de regresión más empleadas son:

- La **regresión lineal** [69] es la más utilizada para formar relaciones entre datos. Es rápida y eficaz pero generalmente insuficiente en espacios multidimensionales donde puedan relacionarse más de dos variables.
- Las **máquinas de vectores de soporte**, en inglés *support vector machines*, son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vapnik y sus colaboradores [70, 71, 72]. Este método es aplicable tanto a problemas de clasificación como de regresión.

- Los **splines o trazadores de regresión adaptativos multivariantes**, denominados en inglés *multivariate adaptive regression splines*. Se trata de una técnica de regresión no paramétrica introducida por [73] y que puede considerarse como una extensión de los modelos lineales que es capaz de modelar no linealidades e interacciones entre variables.

Finalmente, cabe señalar que si bien las redes neuronales artificiales también se pueden emplear como una técnica de regresión más, dado su interés se han introducido en esta relación de forma independiente.

1.9. Aplicaciones del aprendizaje automático a los estudios de genoma amplio

Ya en una fecha tan temprana como 2009, un trabajo publicado por Szymczak et al. [74] adelantaba algunos de los potenciales beneficios de la aplicación de las metodologías de *machine learning* a los estudios de genoma amplio. Así, a lo largo de la ya más de una década transcurrida desde entonces, se han producido múltiples aplicaciones de las distintas metodologías de *machine learning* en estudios de genoma amplio.

Tal y como ya se ha indicado con anterioridad, los algoritmos de aprendizaje automático construyen modelos matemáticos que aprenden a partir de conjuntos de datos. Dentro de los modelos de *machine learning*, éstos, como se indicó anteriormente, se dividen fundamentalmente en aprendizaje supervisado y no supervisado. El aprendizaje supervisado consiste en algoritmos de aprendizaje automático con datos de entrenamiento etiquetados y tiene como objetivo inferir un modelo que sea capaz de relacionar las variables de entrada con las de salida en los modelos de clasificación. Esta función de mapeo puede emplearse para clarificar nuevos datos. Este tipo de modelos suelen ser los de más interés dentro del contexto de los estudios de genoma amplio.

Por otra parte, los modelos no supervisados se caracterizan por no tener una variable de respuesta, sino que el algoritmo busca encontrar patrones dentro del conjunto de datos que se le presenta sin existir un etiquetado previo.

En el contexto de un estudio de genoma amplio, los modelos de *machine learning* proporcionan una base estadística que puede servir como evidencia complementaria a los estudios experimentales [74]. Así, por ejemplo se han aplicado metodologías de *machine learning* para la identificación de *loci* con el fin de incrementar la potencia estadística de los estudios de genoma amplio [75], la detección de interacciones debida a la epitasis [76], mejora de la estimación del riesgo poligénico [77] o bien para la priorización de variantes genéticas en análisis GWAS [78].

Según un artículo de Nichols et al., publicado en 2020 [79], a fecha de septiembre de 2019, el catálogo conocido como NHGRI-EBI GWAS [80] consta de 161.525 asociaciones de variantes genómicas con rasgos, obtenidos a partir de los resultados proporcionados por 4.298 publicaciones distintas. Por tanto, se puede afirmar sin duda que casi desde el mismo comienzo del desarrollo de los estudios de genoma amplio se ha hecho uso de metodologías *machine learning* [81] en este tipo de estudios.

1.9.1. Algunos modelos de aprendizaje automático aplicados a estudios de genoma amplio

La consideración de los estudios de genoma amplio como un problema de clasificación, aconsejan el uso de diversas metodologías de *machine learning* con complejidades muy distintas. La Tabla 1.1 presenta un resumen de algunos estudios efectuados haciendo uso de esta aproximación. Hasta el momento, en los estudios GWAS se ha hecho uso fundamentalmente de cinco metodologías distintas: regresión logística [31], máquinas de vectores de soporte [82], *random forest* [83], *gradient boosting* [84] y redes neuronales profundas [85].

La regresión logística es un método estadístico comúnmente aplicado. Cuando se usa con variables categóricas, puede contemplarse como un modelo lineal generalizado [98]. En una regresión logística, es típico aplicar un término de regularización, por ejemplo, L1 (la suma del valor absoluto de las ponderaciones de las características) y L2 (la suma de las ponderaciones de las características al cuadrado), que introduce algún sesgo al tiempo que reduce la varianza, lo que mejora la capacidad predictiva del modelo obtenido. Los estudios realizados por Demir-Kavuk et al. [94] e Isakov et al. [95] emplearon la regresión logística [99] que combina las penalizaciones L1 y L2 para priorizar los genes que pueden ser responsables de la enfermedad inflamatoria intestinal [100]. Este método realiza una selección de variables (L1) y reduce el tamaño de los coeficientes para reducir la varianza (L2) [32]. La regresión logística regularizada con red elástica tiene como objetivo reducir los problemas relacionados con la conocida como maldición de la dimensionalidad [101], problema muy común dentro de los análisis GWAS dado que el número de SNPs que se emplean normalmente supera con mucho al número de casos y controles disponibles para el estudio. Así por ejemplo, el estudio realizado por Isakov et al. (2017) [95], aplicando regresión logística con red elástica, seleccionó la información más relevante. Dado el tamaño cada vez mayor de los datos genéticos y la gama más amplia de características que están disponibles para describir genes y variantes, la mayor demanda computacional requiere modelos más avanzados.

Siete de los dieciséis trabajos relativos a modelos de *machine learning* aplicados a estudios GWAS que se describen en este apartado, son modelos compuestos; más concretamente del tipo *random forest* y *gradient boosting*. Los métodos compuestos combinan múltiples modelos con el objetivo de mejorar el rendimiento. Estos métodos son de gran utilidad en aquellos estudios de genoma amplio que hacen uso de datos heterogéneos. Deo et al. [88] desarrollaron un modelo de la familia *gradient boosting* denominado «OPEN – Objective Prioritization for Enhanced Novelty» que es capaz de priorizar los genes que están relacionados con algunas enfermedades. Para alimentar este modelo emplearon información proveniente de distintas bases de datos públicas tales como Gene ontology (GO)[102], Mouse Phenotype database [103], Human Phenotype Ontology (HPO) [104] y Online Mendelian Inheritance in Man (OMIM) [105] el uso de toda esta información se beneficia de una reducción de sesgos. El método de *gradient boosting* se basa en modelos de árboles, con tres ramas que toman decisiones del tipo sí/no y que llevan a la clasificacación simple de las muestras [84]. El trabajo de Deo et al. (2014) [88] fue capaz de detectar algunos genes relacionados con enfermedades cardiovasculares. En este caso, el rendimiento de la metodología propuesta se midió haciendo uso del área bajo la curva ROC, obteniéndose valores comprendidos entre 0,75 y 0,9 dependiendo del rasgo objeto de estudio

TABLA 1.1: Referencias bibliográficas relativas a aplicaciones de metodologías de *machine learning* en estudios de genoma amplio.

Descripción del método	Referencia
Modelo *bayesian latent variable*	[86]
Uso de máquina de vectores de soporte y LASSO (*least and shrinkage and selection operator*). Base de datos de 186 pacientes a los que se aplicó la metodología de 80 % de datos para entrenamiento y 20 % para la validación así como *5-fold cross validation*	[12]
Uso de máquina de vectores de soporte	[87, 82, 78]
Uso de árboles de clasificación para la priorización de SNPs relacionados con fenotipos. Se hace uno se seis rondas de *8-fold cross validation*	[88]
Comparación de las metodología de máquinas de vectores de soporte y de *random forest*	[89]
Empleo de una metodología basada en algoritmos evolutivos	[90]
Equilibrado de casos y controles por medio del uso de muestreo. HyperSMURF es capaz de detectar variantes asociadas con enfermedades.	[91]
Uso de redes neuronales profundas, con ayuda de metodología *10-fold cross validation* sobre cuatro conjuntos de entrenamiento y validación diferentes	[92]
Uso de redes neuronales profundas	[85, 78, 93]
Empleo de regresión logística, método del gradiente descendente, máquina de vectores de soporte, *k-nearest neighbor*. Diseño y entrenamiento de una red neuronal configurada para al detección del fenómeno de la epítasis y que es capaz de discriminar distintas características con el fin de poner de manifiesto los genes más relevantes frente al cáncer colorrectal	[87]
Empleo de regresión logística	[94, 95, 87]
Empleo de *random forest*	[78, 95, 87]
Empleo de *extremely randomized trees*	[78]
Empleo de *k-means*	[96]
Empleo de *gradient boosting*	[84, 66, 78]
Regresión lineal multivariante por pasos aplicada a datos genéticos y epigenéticos. Metodología *10-fold cross validation*	[97]

[88]. En este caso, la obtención de un rendimiento tan alto se debe al uso de modelos compuestos lo que proporciona la oportunidad de que los errores de predicción cometidos por uno o unos pocos de los modelos sean corregidos por los resultados obtenidos por la mayoría de los mismos. Nótese también que esta forma de proceder permite ampliar el espacio de búsqueda empleado [106].

Los modelos de *gradient boosting* se caracterizan por su capacidad para reducir el sesgo y la varianza, así como por ofrecer una buena precisión [84]. Sin embargo, también existe la necesidad de comparar el rendimiento del modelo, ya que si bien los modelos de conjunto son robustos, un enfoque singular en un nuevo problema de clasificación proporciona un riesgo de sobreajuste, que también es un problema conocido para el aumento del gradiente dependiendo de las técnicas de regularización utilizadas.

Vitsios y Petrovski [78] construyeron una serie de modelos de aprendizaje semisupervisado, en concreto siete, a través de los que compararon diversas técnicas (*random forest*, *extremely randomized trees*, *gradient boosting machines* GBM, *extreme gradient boosting*, máquinas de vectores de soporte, redes neuronales profundas y un clasificador que hacía uso simultáneo de todos los modelos). Este marco lo emplearon con el fin de priorizar la importancia de algunos genes en tres enfermedades: esclerosis lateral amiotrófica, enfermedad renal crónica y epilepsia. En total, utilizaron datos que contenían más de 1200 características que describen decenas de miles de genes para cada enfermedad. Los resultados obtenidos pusieron de manifiesto que el modelo *random forest* era el clasificador de mejor rendimiento [83]. El método denominado *gradient boosting* fue el segundo más preciso, mostrando el alto rendimiento de la clasificación de conjuntos basada en árboles. Sin embargo, las áreas bajo la curva ROC obtenidos por todos estos algoritmos resultaron ser muy similares como para concluir la existencia de modelo alguno que fuera claramente superior a todos los demás en su aplicación a los conjuntos de datos objeto del estudio. Estos resultados también fueron respaldados por la comparación con el modelo que hacía uso de una combinación de todos ellos, lo que garantizaba la mayor fiabilidad del clasificador elegido para cada enfermedad [78].

El trabajo realizado por Kafaie et al. [87] tuvo como objetivo priorizar los genes asociados con el cáncer colorrectal comparando varios modelos, concretamente, máquinas de vectores de soporte, *random forest*, regresión logística con descenso de gradiente estocástico y k vecinos más cercanos. Los resultados obtenidos llevaron a la conclusión de que la regresión logística era el modelo de mayor rendimiento, poniendo además de manifiesto que, a veces, un problema de clasificación en apariencia complejo puede resolverse de manera óptima con modelos simples.

Además del aprendizaje por medio de modelos agregados y de la regresión logística, las máquinas de vectores de soporte también se han empleado de forma recurrente en estudios que realizan comparaciones de referencia [95, 107, 89, 78]. Las máquinas de vectores de soporte se basan en el cálculo de un hiperplano que es el mejor separa la información correspondiente a los casos de la de los controles [108]. Sin embargo, dentro de los estudios de evaluación comparativa realizados, no ha demostrado ser el modelo de mejor rendimiento. Por ejemplo, en el trabajo realizado por Vitsios y Petrovski [78] se encontró que tenía el área bajo la curva ROC más bajo, 0,83, algo menor que el obtenido por el modelo *random forest* con un área bajo la curva ROC de 0,85. Kafaie et al. [87] encontraron que un modelo de máquinas de

vectores de soporte era capaz de obtener para su problema un rendimiento mejor que el del *random forest*, pero peor que la regresión logística. En lo referente a este rendimiento variable de las máquinas de vectores de soporte según el problema al que se aplican, también debe tenerse en cuenta la importancia de los datos de entrada. Así, el estudio realizado por Kafaie et al. en 2019 [87] ha sido uno de los pocos estudios publicados hasta el momento que se ha centrado en comparar métodos de selección de características y modelos. Así, en este trabajo [87] encontraron que las máquinas de vectores de soporte tuvieron un buen rendimiento, mientras que en comparación, la regresión logística tuvo un rendimiento alto de manera estable, independientemente de la selección externa, enfatizando el valor de la selección de características internas de la regresión logística a través de la regularización.

Dentro del campo de los estudios de genoma amplio, También se ha explorado el aprendizaje profundo para la priorización. Este método puede aumentar la sensibilidad en conjuntos de datos grandes, dada su capacidad para capturar de forma incremental representaciones abstractas de información de alto nivel. En general, esto es beneficioso para la priorización en los estudios GWAS, donde los datos han crecido considerablemente tanto en tamaño como en heterogeneidad y también tiene pocas muestras etiquetadas (variantes / genes causantes de enfermedades conocidas) para el aprendizaje supervisado. Por tanto, el aprendizaje profundo se vuelve ventajoso en este escenario, ya que identifica patrones complejos a través del aprendizaje supervisado y no supervisado de grandes conjuntos de datos [109]. Sin embargo, si bien el aprendizaje profundo permite considerar millones de parámetros, sus aplicaciones fundamentales hasta la fecha se han encontrado lejos del ámbito de los estudios de genoma amplio, aplicándose fundamentalmente a la clasificación de imágenes y al procesamiento del lenguaje natural [110, 111, 112], requiriendo una inversión en su desarrollo y evaluación comparativa con modelos tradicionales para el desarrollo de aplicaciones a estudios de genoma amplio.

En lo referente a aplicaciones ligadas a estudios GWAS, se puede destacar la red neuronal profunda denominada ExPecto [93]. Dicha red fue capaz de priorizar las variantes causales de las enfermedades relacionadas con el sistema inmunológico utilizando características basadas en secuencias. Este conjunto de datos contenía más de 140 millones de mutaciones. Para poder procesar este gran conjunto de datos, ExPecto aplica una metodología basada en la transformación espacial, ponderando las transformaciones en función de las distancias del sitio de inicio de la transcripción. ExPecto también puede realizar el reconocimiento de patrones y la priorización de variantes raras y no observadas. Sin embargo, aunque los modelos se seleccionan en función de su idoneidad para los datos, el rendimiento también puede depender del equilibrio de clases y de la calidad de los datos disponibles.

En general, los estudios de *machine learning* utilizan la validación cruzada para garantizar que los resultados obtenidos sobre una base de datos determinada constituyen una estimación fiable del rendimiento del modelo. Sin embargo, dado que los datos de los estudios de genoma amplio generalmente carecen de variantes y genes causantes de enfermedades validados funcionalmente, existen pocas oportunidades de aprendizaje para los modelos supervisados. Así, resulta posible hacer uso, por ejemplo, de sobremuestreo o submuestreo para abordar el desequilibrio de clases. Schubach et al. [91] desarrollaron un modelo de hiperconjunto (hyperSMURF) utilizando *random forest* con *conciencia de desequilibrio* mediante el uso de submuestreo y

sobremuestreo. Al equilibrar las clases de datos de entrenamiento y exponer el modelo a diferentes conjuntos de datos, los modelos *random forest* pueden diversificar su comprensión de los datos, mejorando la precisión independientemente del tamaño de los datos. Sin embargo, las técnicas de sobremuestreo desarrollan muestras sintéticas basadas en puntos de datos de ejemplo para aumentar el tamaño de la clase minoritaria, lo que puede crear un sobreajuste. Schubach et al. [91] abordaron este problema evitando que se produzcan variantes de ejemplo de la misma ubicación o gen en los conjuntos de entrenamiento y prueba, minimizando de esta forma el sesgo de sobremuestreo.

Finalmente, cabe señalar que, en general, existe la necesidad de realizar evaluaciones comparativas para seleccionar el modelo que mejor se adapta a los datos. En la actualidad, la determinación del modelo óptimo varía entre enfermedades sin un modelo que haya demostrado ser el mejor para todos los casos. Además, la selección del modelo óptimo también parece depender del tamaño y la calidad de los datos. Nótese que los estudios existentes presentan gran variabilidad en lo referente al número de SNPs y pacientes disponibles. Además, los métodos de simulación en ordenador deben tener en cuenta la falta de genes asociados funcionalmente validados, a disposición de los modelos de *machine learning*, y cómo se utilizan las características para construir un modelo adaptado a las características de los estudios de genoma amplio.

Capítulo 2

Hipótesis y objetivos

2.1. Hipótesis

La hipótesis de este proyecto de investigación se enuncia como:

Es posible el desarrollo de una metodología basada en técnicas de *machine learning* capaz de detectar qué SNPs son relevantes para la manifestación de cierto rasgo. Además, dicha metodología es aplicable en el contexto de los estudios de genoma amplio.

Nótese que en este enunciado se considera que los SNPs relevantes provienen de una serie de *pathways* previamente determinados.

2.2. Objetivos

2.2.1. Objetivo general

Tal y como se pone de manifiesto en la hipótesis expuesta más arriba, el objetivo general de este trabajo consiste en el desarrollo de una nueva metodología basada en *machine learning* que sea capaz de determinar qué SNPs pertenecientes a un *pathway* dado son más relevantes a la hora de diferenciar en un estudio de genoma amplio entre casos y controles. Más concretamente, el método que se propone se basa en el uso de un tipo de algoritmos evolutivos denominado algoritmos genéticos, pero que nada tiene que ver más allá de su nombre con la genética biológica, así como de las máquinas de vectores de soporte, metodología propia del *machine learning* muy adecuada para los problemas de clasificación.

Este objetivo general se puede subdividir a su vez en una serie de objetivos específicos.

2.2.2. Objetivo específico número 1

Desarrollo de un algoritmo basado en técnicas de *machine learning* capaz de manejar grandes volúmenes de información y de aplicación en estudios de genoma amplio.

2.2.3. Objetivo específico número 2

Optimización de los parámetros empleados en el algoritmo desarrollado. Para la consecución de este objetivo se considera de interés el uso de la metodología de diseño de experimentos.

2.2.4. Objetivo específico número 3

Comprobación del rendimiento del algoritmo desarrollado aplicándolo a distintos *pathways*. Para poder alcanzar este objetivo, al igual que en el caso del objetivo específico anterior, resulta imprescindible disponer de una base de datos a la que se le pueda aplicar el algoritmo desarrollado.

Capítulo 3

Material y metodología

3.1. Los estudios de genoma amplio

El genoma humano es diploide y comprende aproximadamente tres mil millones de pares de bases, el 3 % de las cuales muestra variación entre individuos según una estimación realizada teniendo en cuenta los 84,7 millones de SNPs analizados en el proyecto de secuenciación del genoma humano [3]. Tanto los trabajos realizados en el mencionado proyecto [3] como los realizados en el marco del proyecto HapMap [13], tuvieron como objetivo genotipar un elevado número de genomas para detectar y anotar las variaciones genéticas entre individuos.

Para realizar un estudio de asociación de genoma amplio, es necesario disponer de dos tipos de información correspondiente a los sujetos del estudio, su fenotipo y su genotipo. Este último puede obtenerse mediante tecnologías de secuenciación de última generación [113], o mediante una matriz de genotipado [114]. Cuando se hace uso de metodologías univariantes, la asociación entre los SNPs y el fenotipo se calcula a través del valor p que se obtiene de la aplicación de un test univariante. Se puede considerar que este valor p representa la probabilidad existente de observar una asociación, bajo la hipótesis nula de no asociación entre el SNP y el fenotipo. Si el valor p cae por debajo de un umbral predefinido, la hipótesis nula es rechazada, lo que significa que existe una asociación entre el SNP y el fenotipo. A pesar de la existencia de una evidencia fuerte contraria a la confirmación de la hipótesis nula, existe una probabilidad de de que la asociación encontrada se debe puramente al azar y que el valor detectado lo hayas sido por pura casualidad, tratándose en ese caso de un falso positivo. Evitar falsos positivos es uno de los principales desafíos a los que se enfrentan los estudios GWAS.

3.1.1. Preprocesamiento de la información y control de calidad

En todo estudio de genoma amplio, antes de proceder a la verificación de las asociaciones entre los SNPs y el fenotipo de interés, es necesario realizar una serie de tareas de preproceso cuyo objetivo es minimizar las posibilidades de encontrar resultados falsos de asociación. Dichas tareas son las que se relacionan y describen a continuación.

- **Transformación de fenotipos** Algunos métodos estadísticos hacen una suposición particular sobre la distribución del fenotipo y sus posibles alteraciones (ruido). Así, por ejemplo, cuando se aplica la regresión lineal en un estudio de genoma amplio, uno de los supuestos del método es que, dado el genotipo, el fenotipo sigue un distribución normal. En la práctica, sin embargo, esta suposición rara

vez se sostiene. Por lo tanto, una tarea común de preprocesamiento es transformar o normalizar los datos fenotípicos de modo que sigan la distribución esperada por el modelo estadístico que se utilice.

- **Filtrado teniendo en cuenta la Ley de Hardy – Weinberg** El equilibrio propuesto por Hardy y Weinberg [115] es un modelo que permite predecir las frecuencias genotípicas de una generación teniendo en cuenta cuáles habían sido dichas frecuencias en la generación anterior. Así, en el contexto de la terminología usual en genética de poblaciones, la ley de Hardy-Weinberg afirma que, bajo ciertas condiciones, tras una generación de apareamiento al azar, las frecuencias de los genotipos de un *locus* individual se fijarán en un valor de equilibrio particular. También especifica que esas frecuencias de equilibrio se pueden representar como una función sencilla de las frecuencias alélicas en ese *locus*. Por tanto, es de esperar que los SNPs objeto de análisis en un estudio GWAS se encuentren en este equilibrio y cualquier desviación pueda evaluarse mediante una prueba estadística con un umbral común en el valor p de 10^{-6} [116]. Nótese que este umbral, aunque no es el único se usa comúnmente en humanos. Si se determina que un SNP no está en en el equilibrio de la Ley de Hardy - Weinberg, esto normalmente se debe a un error de muestreo o genotipado. Cuando esto ocurre, el SNP se elimina del análisis como parte del preprocesamiento de la información.

- **Filtrado por el alelo de menor frecuencia** Se entiende por frecuencia del alelo menos común a la frecuencia del alelo menos común en un determinado locus bialélico [117], dentro de una población. También puede definirse como la frecuencia del segundo alelo más frecuente, en el caso de tener dos o más alelos para ese locus concreto. Este indicador se emplea para estudiar la variación genética, ya que proporciona información que permite diferenciar entre variantes frecuentes y raras en la población. Así, aquellos SNPs que presentan unas frecuencias alélicas bajas, es decir, con valores inferiores a 0,05 o 0,01, se denominan variantes raras y normalmente se excluyen en la mayoría de los estudios GWAS [118]. A no ser que los tamaños de muestra sean muy grandes o bien los efectos de las variantes raras muy evidentes, los análisis de genoma amplio suelen carecer de la potencia suficiente para detectar asociaciones con variantes raras [118, 119]. Existe una familia completa de métodos que son adecuados para el tratamiento de las variantes raras de los alelos [114, 120, 121]. Estos métodos se denominan pruebas de carga. Además de las pruebas de carga, existen otros métodos como la prueba C-alfa [122] o SKAT [119] que han demostrado su eficacia en el análisis de variantes raras. En la actualidad se dispone de bibliografía que explica metodologías adecuadas para el tratamiento de las variantes alélicas consideradas como raras [114, 120, 121, 122, 119].

- **Filtrado de información incompleta** Otro paso importante dentro del proceso de control de calidad que debe realizarse antes de acometer un estudio de genoma amplio, consiste en excluir del mismo tanto a los individuos de los que no se disponga de gran parte de su información genética como de aquellos SNPs para los que no se conozca su alelo en gran parte de los individuos de la muestra [116]. El primer caso es normalmente una consecuencia de la mala calidad del ADN empleado o bien de su baja concentración. En el segundo caso, cuando un SNP tiene una alta tasa de valores faltantes, se considera de baja calidad y se excluye del análisis para evitar falsos positivos [123]. Nótese que

en la actualidad existen protocolos bien definidos para imputar los valores de los SNPs faltantes [124].

- **Corrección de la muestra** Tal y como se mencionó en la introducción, la presencia de grupos de población de características marcadas, puede dar lugar a la aparición de asociaciones fálsamente positivas dentro del conjunto de datos. La forma más frecuente de evaluar el grado de estructura de la población en la base de datos disponible consiste en calcular el factor de inflación genómica [7, 125, 126, 11, 127, 34]. Este factor describe la desviación con respecto a la mediana de la prueba observada.

3.2. La base de datos

La base de datos empleada en este estudio pertenece al proyecto denominado *Colorectal Cancer Transdisciplinary Study* (CORECT) *Project*. Se trata de un proyecto que fue financiado por el National Institute of Health de los Estados Unidos de América a través de la beca número U19CA148107 y que lideró Stephen Gruber, investigador de la *University of Southern California*. En el consorcio creado para este proyecto participaron grupos de investigación con experiencia previa en los estudios de genoma amplio aplicados al cáncer colorectal (WHI, PLCO, CCFR, MECC) por lo que disponían ya de bases de datos y, además, en el marco del mismo, se genotiparon otros 5099 casos y 4830 controles de 6 centros. Así, la base de datos total de este proyecto dispone de unos 10.000 casos y otros tantos controles. A partir de los estudios GWAS, que empleaban diferentes arrays, se ha realizado una imputación de genotipos mediante el software IMPUTE2. Así, se han obtenido datos de más de 38 millones de SNPs para todos los individuos. Como muchos de estos SNPs son de frecuencia muy baja, se han depurado los datos, eliminando los SNPs con baja calidad de imputación y aquellos con frecuencia alélica inferior al 1 %. De la mayoría de los individuos incluidos en estos estudios se dispone de información epidemiológica detallada y se ha realizado el trabajo para armonizar las variables importantes para el cáncer colorectal.

Este estudio multicéntrico y observacional se llevó a cabo desde septiembre de 2008 hasta diciembre de 2013. Para el trabajo que se presenta en este proyecto de investigación, el conjunto de información que se ha empleado es el correspondiente tanto al Hospital Universitario de León como al Hospital de Bellvitge. Dicho conjunto contiene información relativa a 1076 casos y 973 controles de cáncer colorectal, de los que se hizo uso de la información correspondiente a 370.570 SNPs diferentes.

Los casos incluídos en la base de datos fueron previamente confirmados por medio de histología. Todos los pacientes del estudio presentaban edades comprendidas entre los 20 y los 85 años. La existencia de dificultades para la comunicación del paciente, así como el padecimiento previo de cáncer colorrectal, fueron criterios de exclusión del estudio.

Los controles participantes en el estudio se eligieron de forma aleatoria de entre los pacientes asignados a los médicos de atención primaria pertenecientes a las zonas sanitarias de cada uno de estos hospitales. Así, se eligieron controles del mismo sexo y rango de edad con una diferencia máxima de 5 años con respecto a los

TABLA 3.1: *Pathways* analizados en el presente proyecto de investigación.

Pathway	Referencias
Adipocytokine signaling pathway	[131, 132, 133]
AMPK signaling pathway	[134, 135, 136]
Apelin signalling pathway	[137, 138, 139]
Colorectal cancer pathway	[140, 141]
Glucagon signalling pathway	[142, 143]
Enfermedad de Huntington	[144, 145]
Insulin resistance	[146, 147]
Insulin signalling pathway	[148, 149, 150]
Longevity regulating pathway	[151, 152, 153]
Biogénesis mitocondrial	[96, 154]

pacientes con cáncer colorectal. Todos los controles selecionados debían de llevar residiendo en la zona sanitaria al menos 6 meses.

Los protocolos del estudio fueron aprobados por los comités de ética de las instituciones que formaron parte del mismo. La participación de todas las personas reclutadas fue voluntaria, y se produjo tras la firma de un consentimiento informado. La confidencialidad de los datos se garantizó por medio de la eliminación de los datos de carácter personal. Todos los ficheros de datos empleados en este estudio cumplen con la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Además, dichos ficheros se registraron en la Agencia Española de Protección de Datos con el número 2102672171.

3.2.1. Pathways analizados

En el presente proyecto de investigación se han considerado diez *pathways* distintos, todos ellos pertenecientes a la base de datos KEGG (Kyoto Encyclopedia of Genes and Genomes) [128, 129, 130]. KEGG es una base de datos de gran utilidad para comprender las funciones y utilidades de alto nivel del sistema biológico, a partir de información molecular, especialmente de conjuntos de datos moleculares a gran escala generados por secuenciación del genoma. La lista de los pathways considerados se puede consultar en la Tabla 3.1. La selección de estos *pathways* se hizo en función de sus características, y por eso, algunos de los escogidos los fueron por tener, según la bibliografía, una relación clara y conocida entre ellos con el rasgo objeto de estudio (padecimiento de cáncer colorrectal), mientras que otros se eligieron por tener una relación probable pero no confirmada, siendo dicha asociación dudosa y, finalmente, otros por tener una relación altamente improbable con el rasgo objeto de estudio, siendo dicha asociación considerada improbable. De esta forma se pretendía disponer de respuestas distintas a la aplicación del algoritmo.

Nótese que, de forma simplificada, es posible afirmar que la metodología general que se encuentra detrás del análisis de *pathways* consiste en la reducción y agregación de datos. En primer lugar, la información relativa al genoma se secuencia y se presenta al investigador en forma de SNPs. Estos SNPs son analizados teniendo en cuenta las diferencias en las frecuencias alélicas existentes entre casos y controles. Además, los SNPs, teniendo en cuenta su posición, se relacionan con los genes y los distintos genes son incluidos en los *pathways* a los que pertenecen.

3.3. Diseño de experimentos

El diseño de experimentos, también conocido por sus siglas en inglés DOE (*design of experiments*), es un enfoque sistemático para el análisis de las relaciones existentes entre una variable dependiente y un conjunto de variables independientes [155]. Por tanto, el diseño de experimentos es una metodología matemática que en la práctica presenta numerosas aplicaciones enfocadas a la optimización del rendimiento de un sistema, maximizando sus propiedades, representadas a través de una o varias variables dependientes [156]. El diseño de experimentos emplea herramientas propias de la estadística, con el fin de analizar los datos y predecir el rendimiento de un sistema dentro de los límites seleccionados para el experimento. Además de ayudar a la comprensión de cómo los cambios en cualquier variable independiente afectan al rendimiento de la variable dependiente, los diseños de experimentos sirven para estudiar las interacciones entre las diferentes variables del sistema objeto de estudio. Por tanto, se puede afirmar que el diseño de experimentos es una técnica o procedimiento que permite analizar las interacciones entre las variables de un sistema haciendo uso del mínimo de información necesaria y empleando [157]:

- Ciertos valores que determinan los límites experimentales.
- Unas condiciones experimentales específicas.
- Información relativa a las variables independientes y a la variable dependiente en los puntos del sistema empleados como referencia para el análisis.

Así, a través de los resultados de un diseño de experimento resulta posible predecir la respuesta del sistema en cualquier punto dentro de los límites experimentales impuestos [156].

El diseño de experimentos se emplea, fundamentalmente, para determinar qué factores o variables e interacciones son importantes y contribuyen de manera determinante al efecto que se mide [157]. El uso del diseño de experimentos, dado que se trata de una aproximación sistemática al estudio de un problema, ahorra tiempo y recursos, proporcionando una comprensión útil de la propiedades y del proceso objeto de análisis. El momento más adecuado para hacer uso del diseño de experimentos es durante el desarrollo de un nuevo producto o proceso así como al resolver los problemas técnicos surgidos en el mismo y en los que intervengan más de una variable [157]. En el caso del presente proyecto de investigación, el diseño de experimentos se ha empleado para optimizar los valores de algunos de los parámetros que intervienen en el algoritmo desarrollado.

Por tanto, el diseño de experimentos resulta de gran utilidad para resolver cualquier problema técnico cuando se necesita entender completamente la respuesta a

diferentes variables de proceso o producto que se puede cambiar o controlar durante la experimentación [158]. La ventaja fundamental de hacer uso de un enfoque basado en diseño de experimentos, se encuentra en que los datos generados de una forma sistemática con la ayuda del diseño de experimentos constituyen el menor conjunto de información necesaria para una correcta evaluación del experimento objeto de estudio. Es decir, los resultados de un diseño de experimentos se caracterizan por proporcionan una comprensión global del proceso, así como de las interacciones de las variables que se producen dentro del espacio experimental objeto de estudio.

Dentro del diseño de experimentos existen tres tipos fundamentales. Estos son el diseño factorial [159], la metodología de superficie de respuesta [160] y los experimentos de mezcla [161]. En el caso del presente proyecto de investigación se ha hecho uso de un diseño factorial.

3.3.1. Las etapas del diseño de experimentos

El proceso necesario para la realización de un diseño de experimentos se puede dividir en cinco etapas fundamentales, las cuales se describen a continuación [162, 163, 164].

3.3.2. La definición del problema

Aunque esta primera etapa pueda parecer obvia, en la práctica, a veces es difícil de hacer. En el caso de un experimento algorítmico como el que se describe en el presente proyecto de investigación, el trabajo de análisis ha sido fundamentalmente realizado por el doctorando, pero en general, cuando se analiza un problema al que se quiere aplicar diseño de experimentos, es frecuente que las aportaciones de ideas por parte de diferentes personas puedan entrar en conflicto. Esto motiva que la decisión de qué variables específicas se deben controlar pueda ser controvertida. Así, incluso puede suceder que en la práctica, algunas opiniones puedan enfatizar la necesidad de hacer uso de la metodología del diseño de experimentos, mientras que otras sugieran que el tiempo y recursos disponibles sería mejor invertirlos en otra cosa. Además, en la etapa de definición del problema también resulta crítico verificar que el problema está correctamente definido.

3.3.3. La planificación del experimento

Una vez que el problema se ha definido correctamente, el segundo paso en el proceso de un diseño de experimentos consiste en la selección de las variables independientes, así como sus límites de evaluación, y la variable o variables dependientes. Se entiende como variable dependiente la respuesta que se medirá en cada uno de los experimentos, que además es la variable cuyo resultado se pretende optimizar. Las variables o factores independientes son parámetros de procesamiento o producto que se establecen en valores específicos (niveles) y se encuentran controlados en el diseño experimental. Las variables o factores dependientes son las respuestas que se miden para cada experimento con el fin de determinar si las variables independientes tienen un efecto sobre las propiedades o las condiciones de procesamiento evaluado.

Las variables independientes son normalmente de naturaleza cuantitativa, lo que significa que se establecen en un valor numérico. Sin embargo, en algunos diseños

de experimentos, las variables independientes son cualitativas. En el caso del diseño de experimento realizado en la presente investigación, se hizo uso de variables cuantitativas y cualitativas, pero en general, siempre que sea posible, resulta más sencillo hacer uso únicamente de factores cuantitativos[164].

El número de puntos de datos evaluados para cada respuesta experimental, dependerá del número de pruebas requeridas para la significanción estadística, basadas en la precisión y exactitud de la prueba. La precisión está relacionada con cómo de cerca se encuentran las medidas entre sí. Si la precisión de la medición es alta, el número de muestras necesarias para obtener una media precisa es menor que si la precisión es baja. Téngase en cuenta que la precisión mide la distancia entre los valores de un test y las medidas reales. En la práctica, resulta posible tener alta precisión y baja exactitud.

Finalmente, cabe señalar que los experimentos que componen cualquier aplicación de la metodología de diseño de experimentos, se deben realizan en orden aleatorio para minimizar el posible error sistemático existente, dado que este podría conducir a conclusiones erróneas.

3.3.4. La recopilación de datos

El tercer paso del proceso necesario para la realización de un diseño de experimentos es la recopilación de datos. Una vez definidos los experimentos, llega el momento de ejecutar el algoritmo para los distintos valores de las variables objeto de estudio. Tal y como ya se ha indicado, esta ejecución se debe realizar en orden aleatorio.

3.3.5. El análisis de datos

El cuarto paso en el proceso dentro de un diseño de experimentos es analizar los datos. Este análisis se realiza normalmente con la ayuda de un soporte informático. En el caso del presente proyecto de investigación se ha hecho uso del programa Minitab (Minitab LLC; Pine Hall Road, State College, Pensilvania, Estados Unidos). Gracias al uso de este programa, resulta inmediato conocer qué factores e interacciones son importantes para una respuesta particular, generar modelos para predecir la respuesta dependiente en cualquier punto experimental, representar gráficamente las ecuaciones del modelo para proporcionar una comparación visual de los datos, predecir la condición experimental del proceso donde la respuesta es máxima, y predecir o definir un rango de operación experimental donde las propiedades cumplan con las especificaciones o valores deseados.

Así, las conclusiones que se obtengan se basarán en el análisis estadístico y en los posibles niveles de confianza. Las estadísticas no pueden probar que un factor realmente tiene un efecto sobre la variable de salida, pero aportan confiabilidad estadística y prueban la validez de las hipótesis. Así, el establecimiento de un límite de confianza determina si factor en particular presenta algún efecto, basado en cierto nivel de confianza. Normalmente, se emplea un nivel del 95 %.

TABLA 3.2: Efectos de los coeficientes en el ejemplo propuesto.

	Efectos de los coeficientes			
	a_1b_1	a_2b_1	a_1b_2	a_2b_2
A	-1	+1	-1	+1
B	-1	-1	+1	+1
AB	+1	-1	-1	+1

3.3.6. La presentación de las conclusiones

El quinto y último paso de todo diseño de experimentos consiste en la presentación de las conclusiones. En el caso del presente proyecto de investigación, los resultados obtenidos se presentan dentro de la sección del mismo título de este documento.

3.3.7. Formulación matemática de un diseño factorial

Se introduce a continuación la formulación matemática del diseño de experimentos factorial [165, 166, 167], pues se trata de la metodología de diseño de experimentos que se ha empleado en el presente proyecto de investigación.

Dado un experimento en el que intervienen dos variables independientes A y B, cada una de ellas con dos niveles, un diseño de experimentos *ad hoc* debería de medir el efecto de las combinaciones (a_1, b_1), (a_1, b_2), (a_2, b_1) y (a_2b_2). Dado que se dispone de dos variables, cada una de ellas con dos niveles, este diseño factorial sería 2^2. Normalmente, en un diseño factorial se hace uso de n variables, cada una de las cuales tiene 2 niveles. Por tanto, la forma general de un diseño factorial es 2^n.

Con el fin de conocer el efecto de A, se emplearía la ecuación:

$$A = (a_2 \cdot b_1 - a_1 \cdot b_1) + (a_2 \cdot b_2 - a_1 \cdot b_2)$$

De manera análoga, para conocer el efecto de B se usaría:

$$B = (b_2 \cdot a_1 - b_1 \cdot a_1) + (b_2 \cdot a_2 - b_1 \cdot a_2)$$

Si se lleva a cabo un experimento tradicional, se debe modificar el valor de cada variable de manera separada. Esta forma de proceder habría dado lugar a la necesidad de realizar ocho experimentos. Por tanto, se comprueba cómo el uso del diseño de experimentos ahorra tiempo y dinero en el proceso de análisis. Así, el planteamiento del diseño de experimentos con dos variables al que se ha hecho referencia es el que se muestra en la Tabla 3.2.

Es necesario indicar que AB se calcula multiplicando a_x por b_x con el fin de encontrar el efecto combinado de los coeficientes. Una condición adicional que debe tenerse en cuenta es que, en general, será necesario hacer uso de más de una réplica

de cada uno de los experimentos con el fin de conseguir un resultado lo suficientemente preciso. Para calcular el efecto medio de un factor, el valor obtenido se divide por 2 veces el número total de réplicas efectuadas. Nótese que se llama réplica a cada una de las repeticiones del experimento. Por tanto, se hará uso de la ecuación:

$$media\ del\ efecto\ del\ factor = \frac{efecto\ factorial\ total}{2 \cdot numero\ replicas}$$

Si se añade una tercera variable C, el proceso de obtener los coeficientes se vuelve significativamente más complicado. Así, el efecto de A vendría dado por la ecuación:

$$A = (a_2 \cdot b_1 \cdot c_1 - a_1 \cdot b_1 \cdot c_1) + (a_2 \cdot b_2 \cdot c_1 - a_1 \cdot b_2 \cdot c_1) + (a_2 \cdot b_1 \cdot c_2 - a_1 \cdot b_1 \cdot c_2) + (a_2 \cdot b_2 \cdot c_2 - a_1 \cdot b_2 \cdot c_2)$$

La tabla de coeficientes que se generaría, sería también notablemente más compleja, lo que hace recomendable el uso de algoritmos de simplificación como el de Yates [168, 169]. Afortunadamente, parte de esta complejidad metodológica es asumida por el software empleado para el diseño de experimentos.

3.4. Los algoritmos evolutivos

Los algoritmos evolutivos son métodos de búsqueda inspirados en los principios de la selección natural y la genética. Existen varios tipos de algoritmos evolutivos. Los que primero se desarrollaron son los conocidos como algoritmos genéticos [63]. Todos los algoritmos evolutivos se basan en alternar mecanismos de selección y variación. Los mecanismos de selección se encargan de concentrar la búsqueda en las áreas que parecen prometedoras, mientras que los mecanismos de variación son los responsables de producir nuevas soluciones a partir de aquellas que se seleccionaron. Las diferencias entre los distintos algoritmos evolutivos tienen orígenes históricos y, comúnmente, se explican a partir de la forma en la que se representan las soluciones. A continuación, se detallan los principales componentes de todo algoritmo evolutivo.

Dado que en el presente proyecto de investigación se presenta una metodología basada en machine learning de aplicación en estudios de genoma amplio, resulta necesario aclarar que, si bien los fundamentos de los algoritmos genéticos están inspirados en la genética biológica y en la evolución natural, estos son una simplificación desarrollada al margen de la genética y con fundamentaciones matemáticas. Por tanto, no se debe buscar que las metodologías de algoritmos evolutivos y más concretamente la de algoritmos genéticos respondan a todos los principios de la genética biológica, sino considerarlos como un desarrollo paralelo inspirado en esta que ha demostrado, desde hace décadas, su utilidad en los campos más diversos [170, 171, 172, 173, 174].

Finalmente, cabe también señalar que los nombres que se emplean para estas metododologías en el presente proyecto de investigación, «algoritmos genéticos» y «algoritmos evolutivos», son los comúnmente empleados en la literatura sobre este tema en lengua española [175, 176, 177, 178, 179, 180].

3.4.1. Función objetivo

La mayoría de las aplicaciones de los algoritmos genéticos son problemas de optimización. La función objetivo representa la función que se desea optimizar. El algoritmo genético usa la función objetivo para asignar a cada individuo de la población un valor de aptitud que será utilizado posteriormente por el mecanismo de selección para identificar las soluciones más prometedoras. Así, la función objetivo se representa como $f : Dom \rightarrow R$, donde Dom es el dominio de la función y el resultado es un número real, pero los algoritmos evolutivos no requieren que la función se pueda expresar de forma simbólica. Nótese que los algoritmos evolutivos funcionan bien en situaciones en que la función objetivo es estocástica, esto es, en aquellas situaciones en las que la función objetivo puede devolver diferentes resultados utilizando las mismas entradas.

3.4.2. Selección

El proceso de selección se encarga de dirigir el algoritmo hacia aquellas regiones del espacio de búsqueda que parezcan prometedoras. Existen diversos mecanismos de selección, pero todos tienden a escoger a los mejores individuos según la función objetivo y a descartar a los peores [181]. La mayoría de los mecanismos de selección son estocásticos, pero también existen versiones deterministas. Por ejemplo, la selección por truncamiento escoge a todos los individuos que exceden de cierto valor de aptitud o que están clasificados entre los primeros m elementos de la población. El método de selección que se relaciona más frecuentemente con los algoritmos genéticos es la selección proporcional [182]. Este método consiste en asignar a cada individuo i una probabilidad de ser seleccionado p_i, de acuerdo con la razón de su valor de aptitud entre la suma de todos los valores de aptitud: $p_i = f_i / \sum_j f_j$. Después se muestra la población utilizando estas probabilidades. Los individuos con mayores aptitudes se seleccionarán más probablemente que los menos aptos.

El método proporcional tiene varias desventajas. Una de ellas es que el algoritmo se comporta de manera diferente si la función objetivo es modificada, por ejemplo, sumándole una constante a la función $f'(x) = f(x) + c$. Al optimizar f', el denominador de las probabilidades p_i es distinto a lo que sería optimizando f. Otra desventaja es que conforme la ejecución del algoritmo avanza, la población tiende a estar compuesta de individuos con características similares. Cuando las características son similares, las probabilidades de selección tienden a ser uniformes. Así, las probabilidades de seleccionar a los mejores individuos son solo un poco mayores que las probabilidades de seleccionar a los peores. Esto ocasiona que el algoritmo genético no progrese rápidamente [183].

Para disminuir este problema de estancamiento, se han propuesto varias alternativas. Una alternativa es escalar los valores de aptitud de cada individuo y utilizar los valores escalados para determinar las probabilidades de selección. Otra alternativa consiste en ordenar los individuos de acuerdo con su aptitud y basar la selección en este orden [182]. Así se evita que individuos muy aptos dominen la población rápidamente, reduciendo la diversidad que es necesaria para explorar más soluciones. Utilizando el rango para seleccionar también se evita que la búsqueda se estanque en las etapas avanzadas.

Una variación que se puede incorporar a muchos métodos de selección es el elitismo [184]. En este tipo de variación se identifica a cierto número de los mejores individuos de una determinada generación, y estos individuos se insertan en la población después de aplicar un método de selección estocástico. Esto asegura que las mejores soluciones encontradas no se pierdan porque tengan la mala suerte de no ser seleccionadas. Generalmente, se elige un número pequeño de los mejores individuos. También es posible insertar individuos después de realizar las operaciones de variación para asegurar que los mejores individuos no se pierdan porque fueron destruidos por estos operadores.

3.4.3. Representación

La forma en la que se representan las soluciones en los individuos, determina en gran manera el éxito de los algoritmos evolutivos. En la práctica, el usuario debe decidir la representación más adecuada al problema que se desea resolver. En algunos casos, la decisión acerca de la representación es sencilla. Por ejemplo, en problemas en los que se debe elegir un subconjunto de variables a partir de un conjunto finito conviene utilizar una representación binaria. En este caso, la selección de cada objeto se puede representar con un 0 o un 1. Así, un subconjunto de objetos está representado por una cadena binaria. Nótese que este tipo de representación es el que se ha empleado en la metodología basada en algoritmos genéticos que se emplea en el presente proyecto de investigación.

3.4.4. Mutación

El método más sencillo para producir nuevos miembros de una población es la mutación [185]. En representaciones binarias este operador identifica aleatoriamente los bits que van a mutar y cambia sus valores de 0 a 1 o viceversa. En los algoritmos genéticos, generalmente, se asigna una probabilidad pequeña para aplicar este operador a cada posición de cada individuo. Históricamente, esta probabilidad se situaba entre 0.01 y 0.001 independientemente del problema, pero más recientemente se ha vuelto común utilizar una probabilidad de $p_m = 1/l$, donde l es la longitud, número de bits, de los individuos [150]. Una forma para implantar este método de mutación es generar un número aleatorio entre 0 y 1 para cada posición de cada individuo. En aquellas posiciones donde el número aleatorio sea menor que p_m, se cambia el valor.

3.4.5. Cruzamiento

El operador de cruzamiento [186] consiste en elegir al azar un punto de cruce en el interior de los cromosomas de dos individuos y formar otros dos nuevos individuos uniendo el segmento izquierdo del cromosoma de un padre con el segmento derecho del otro. El cruce en un punto se puede generalizar a cruce en n puntos, donde los donantes se parten en $n + 1$ segmentos al azar y los descendientes se obtienen tomando segmentos alternos de cada donador. Por ejemplo, en un cruce con dos puntos se obtienen tres segmentos de un donador X que podemos designar como X_1, X_2 y X_3 y otros tres segmentos del donador Y, Y_1, Y_2 y Y_3. Los dos individuos que resultan del cruce en dos puntos son X_1, Y_2 y X_3 e Y_1, X_2 y Y_3.

Otro operador de cruce utilizado frecuentemente en la práctica es el cruce uniforme [187]. Este operador trata cada posición de los descendientes independientemente y decide aleatoriamente el donador para cada posición de los hijos. Esto es, en lugar de dividir los cromosomas en segmentos que se intercambian, para cada posición se elige un donador al azar. Generalmente, se utiliza una probabilidad igual para cada padre (50 %) y se utiliza la misma probabilidad para cada posición, aunque es posible pensar en mecanismos para adaptar las probabilidades [188]. El cruce en n puntos tiende a mantener juntos los genes que están cercanos entre sí en la representación. En cambio, el cruce uniforme ignora completamente la posición de las variables en la representación. Al igual que el operador de mutación, el operador de cruce tiene asociado una probabilidad de cruce. Históricamente, en los algoritmos genéticos, esta probabilidad es relativamente alta, entre 0.6 y 1, en comparación con las probabilidades de mutación. En otros algoritmos evolutivos también se utilizan frecuentemente probabilidades relativamente altas de cruce.

3.4.6. La formulación matemática de los algoritmos genéticos

El teorema del esquema. A lo largo de las últimas décadas, la teoría de los algoritmos genéticos ha sufrido una revisión que ha modificado el planteamiento original propuesto por Holland [189, 190]. A pesar de esto, resulta posible afirmar que los conceptos básicos de la metodología siguen plenamente vigentes. Una de las discusiones clásicas dentro del contexto de la teoría de los algoritmos genéticos, es la conveniencia o no de hacer uso de codificación binaria. Así, el uso de es tipo de codificación lleva a la creación de un espacio de búsqueda específico para este tipo de problemas que se puede nombrar como A^l.

Los Schemata Resulta también conveniente introducir el concepto de esquema o *schema*. Esta palabra viene del tiempo pasado del verbo griego $\epsilon\chi\omega$, que significa tener. Su plural es *schemata*.

Definición. Un esquema es un subconjunto del espacio A^l en el que todas las cadenas comparten una serie definida de valores particulares.

Esto se puede representar por medio de $A \cup *$, donde el símbolo $*$ representa cualquier posible término al que se le aplique la operación. Así, en el caso binario, $(1 * *1)$, representa un hipercubo de cuatro dimensiones $\{0,1\}^4$ en el que tanto el primer como el último gen toman el valor 1. Es decir, representa de forma simplificada a las cadenas $\{(1001), (1011), (1101), (1111)\}$. Por tanto, se puede decir que los schemata son conjuntos de términos teóricos.

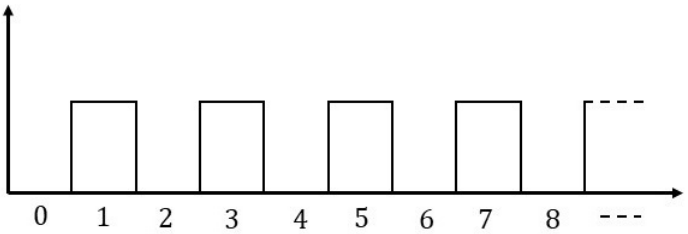


FIGURA 3.1: Posibles valores enteros del esquema $(* \cdots *1)$.

Tal y como se observa en la Figura 3.1 el esquema $(* \cdots *1)$ representa los enteros impares, en un ejemplo en el que lo que se codifican son valores enteros.

En virtud de lo expuesto anteriormente, se puede afirmar que cualquier elemento del espacio muestral considerado no será más que una instancia de muchos posibles schematas [191, 192]. En general, se puede afirmar que si la cadena tiene una longitud l, cada uno de los elementos de la población es una instancia de $|A|^l$ schematas distintos. Como consecuencia de esto, cada vez que se aplica la función de evaluación a un elemento de la población diferente, se obtiene información acerca del valor medio de ajuste de cada schemata del que dicho elemento representa una instancia. Así, en el caso binario, con un tamaño de la población de N se podría disponer de $N2^l$ schemata, pero en la práctica se produce solape entre ellos. Además, la representación de los mismos no será equilibrada, dado que el algoritmo genético se centrará en los que proporcionen un mejor ajuste.

Paralelismo intrínseco o implícito

En sus trabajos de fundamentación de los algoritmos genéticos, Holland [63, 193] afirmaba que, dado que se probaba por medio de un conjunto de espacio muestral formado por un número elevado de elementos y haciendo uso únicamente de un reducido número de los mismos, los algoritmos genéticos poseían una propiedad denominada *paralelismo intrínseco*.

La respuesta a cuántos schemata se procesan en el cálculo de un método de algoritmo genético es, en general, difícil de responder. Un enfoque adecuado para obtener una respuesta a esta pregunta puede consistir en centrarse en aquellos schemata que tienen una cierta probabilidad de sobrevivir $1 - \eta$ [194].

También resulta necesario hacer uso de las nociones de *longitud* y *orden* de un esquema. Por *longitud* se entiende la distancia entre el comienzo y el final de la cadena que constituye cada uno de los elementos de la población, y por orden el número de posiciones definidas. Así por ejemplo, el esquema $(1 * *1)$ tiene una longitud de 3 y su orden es 2.

El valor de η se puede considerar como un cierto «error de transcripción» que se produce en la práctica. Con el fin de estimar el número de schemata que satisfacen estas condiciones, se considerará una ventana $2m$, de elementos contiguos, contados desde la posición más a la izquierda. Así, existen $\binom{2m}{m}$ posibilidades de definir ese subconjunto de elementos para un esquema de orden m en esta ventana, lo que podría dar lugar a 2^m schemata. Así, por ejemplo, $(f * f * \cdots *)$, donde f es un valor fijo, ya bien sea este 0 o 1, representa 4 schemata correspondienes a las 4 formas de elegir los valores fijos. Ahora, de forma sucesiva, esta ventana se irá moviendo una posición hacia la derecha, observando que de cada vez habrá $\binom{2m-1}{m-1}$ conjuntos de posiciones que no han sido definidas en la ventana previa. Téngase además en cuenta [195] que se puede mover esta ventaja un paso hacia la derecha $(l - 2m)$ veces, y que por tanto existen

$$\binom{2m}{m} + (l - 2m)\binom{2m-1}{m-1}2^m = (l - 2m)\binom{2m}{m}2^{m-1} \tag{3.1}$$

schemata posibles para cada cadena $\approx 2^{3m-1}/\sqrt{\pi m}$, si se emplea la aproximación de Stirling para los números factoriales [196].

Llegados a este punto, resulta obvio indicar que en una cadena dada, los 2^m schemata no se encuentran representados por una elección particular de un elemento definido, por tanto, resulta necesario estimar cúantos schemata pueden representar condiciones que ocurran realmente en una población de tamaño N. Si se considera que los miembros de la población se escogen de forma aleatoria haciendo uso de un muestreo aleatorio uniforme, una población de tamaño $N = \nu 2^m$ donde ν representa un entero de valor pequeño, podría suponerse que debería tener ν instancias del schemata de orden m, así que con esta definición de N, se puede estimar que el número de schemata de longitud igual menor que 2^m y orden m, sería $O\left(N^3\right)$. En realidad, se trata de un valor que se encontraría infraestimado, dado que existen muchos otros schemata que no han sido tenidos en cuenta y este valor ha de considerarse únicamente como un límite inferior [197, 198].

El Teorema de los Schemata

Siguiendo la línea re razonamiento establecida, resulta necesario plantearse cómo de relevante es la idea del paralelismo implícito. Desde el punto de vista de cualquier problema real, y más de uno con el tamaño del que se aborda en el presente proyecto de investigación, resulta claramente imposible almacenar los valores promedio estimados por la función objetivo de todos estos schemata de manera explícita, y además, tampoco es obvio cómo explotar dicha información, incluso si fuera posible su almacenamiento. Holland [63, 199] demostró que mediante la aplicación de operadores genéticos como el cruce y la mutación, cada esquema representado en la población actual aumentará o disminuirá de acuerdo con su idoneidad relativa, independientemente de lo que suceda con otros schemata.

Por tanto, lo que Holland probó desde el punto de vista matemático es lo que hoy en día se conoce como el Teorema de los Schemata [63, 199]. Este teorema resulta de gran aplicación dentro del campo de los algoritmos genéticos [200, 201, 202, 203] y se enuncia a continuación sin demostración alguna, pero enunciando previamente algunos lemas que resultan de utilidad para la comprensión del mismo.

Se define la ratio de ajuste $(S,t) = f\left(S,t\right)/\overline{f}\left(t\right)$ como el parámetro que expresa el valor medio de la función de ajuste u optimización de un esquema en relación con el valor de dicha función sobre el conjunto de la población. Se asume por tanto que el ajuste de un esquema es el valor medio que toma la función de ajuste sobre todas las instancias que presenta dicho esquema en la población $P(t)$, es decir,

$$f\left(S,t\right) = \frac{\sum_{x \in S \cap P(t)} f(x)}{|S \cap P(t)|}, \tag{3.2}$$

siendo

$$\overline{f}\left(t\right) = \sum_{x \in P(t)} f(x)/\left|P(t)\right|. \tag{3.3}$$

Nótese que en la función $P(t)$ es posible que se produzcan múltiples ocurrencias de un mismo elemento de la población. Seguidamente, se introduce un lema que es de utilidad para la compresión del Teorema.

Lema 1. Si se considera un plan reproductivo en el que el progenitor se selecciona teniendo en cuenta su función de ajuste, el número esperado de instancias de un esquema S en un momento dado $t+1$ viene dado por

$$E\left[N\left(S,t+1\right)\right] = r\left(S,t\right)N\left(S,t\right) \tag{3.4}$$

donde $N\left(S,t\right)$ representa el número de instancias de S en el momento t.

Lema 2. Si se realiza un cruzamiento en un punto cierto momento temporal t con probabilidad χ en un esquema, S de longitud $l(S)$, la probabilidad de que S sea representado en la población en el tiempo $t+1$ viene acotada inferiormente por

$$1 - \chi\frac{l(S)}{l-1}P_{diff}\left(S,t\right) \tag{3.5}$$

Donde l representa la longitud de la cadena considerada, y $P_{diff}\left(S,t\right)$ es la probabilidad de que el segundo progenitor sea una instancia proveniente de un esquema distinto.

Lema 3. Si además se aplica una operación de mutación en el mismo tiempo t con probabilidad μ para cada elemento de un esquema S de orden $k(S)$, entonces la probabilidad de que S se vea representado en la población en el momento $t+1$ viene acotada inferiormente por

$$1 - \mu k(S) \tag{3.6}$$

La combinación de todos los resultados expuestos anteriormente, conduce a la obtención del siguiente teorema:
Teorma del Esquema. Haciendo uso de un plan reproductivo como el definido anteriormente, en el que las probabilidades de cruzamiento y mutación son de χ y μ respectivamente, y del schema S de orden $k(S)$, y longitud $l(S)$ tiene una ratio de ajuste $r\left(S,t\right)$ en el tiempo t, por tanto, el número esperado de representantes del esquema S en el tiempo $t+1$ viene dado por

$$E\left[N\left(S,t+1\right)\right] \geq \left\{1 - \chi\frac{l(S)}{l-1}P_{diff}\left(S,t\right) - \mu k(S)\right\} r\left(S,t\right)N\left(S,t\right)$$

Nótese también que resulta posible prescindir del término $P_{diff}\left(S,t\right)$ y dejar el Teorema del Schema de la siguiente forma:

$$E\left[N\left(S,t+1\right)\right] \geq \left\{1-\chi\frac{l(S)}{l-1}-\mu k(S)\right\} r\left(S,t\right) N\left(S,t\right)$$

Así la formulación resulta más simple, pero debe tenerse en cuenta que esta no es la única formulación posible para este teorema. Finalmente, cabe también señalar que en este desarrollo se ha asumido que existe un único punto de cruzamiento para la creación de los elementos de la siguiente generación.

Críticas a la formulación matemática de los algoritmos genéticos a través del Teorema del Esquema

La teoría basada en los schemata que propuso Holland ha sufrido críticas por parte de diversos autores, como, por ejemplo, Mühlenbein [204], quien ya en 1991 expuso las limitaciones que en su opinión presentaba una teoría basada en principios evolutivos darwinianos. Concretamente, este autor afirmaba que [204]:

...el Teorema del Esquema es prácticamente una tautología, pues únicamente describe selecciones proporcionales...

Sin embargo, en opinión de otros autores como Reeves et al. [205], esta crítica resulta injusta, pues si el teorema es incorrecto, deberían de presentarse casos en los que no se cumpla y hasta el momento esto no ha sido así. Además, en cierta medida, todo teorema matemático es en sí mismo una tautología pues sus conclusiones se encuentran de forma inherente en sus premisas. Realmente, a lo que Mühlenbein se refería es a que el conocimiento que aporta este teorema es aparentemente trivial para el esfuerzo de formalización que supone.

Sin embargo, otros autores, entre los que se encuentran Radcliffe [206] y Vose [205, 207] afirman que es posible extender el Teorema del Esquema a un subespacio arbitrario del espacio de búsqueda. Así por ejemplo, los resultados obtenidos por Vose [205] mostraron cómo un pequeño cambio en la tasa de mutación de un algoritmo genético puede dar lugar a un cambio de gran importancia en la evolución de las soluciones del mismo, tratándose además de un cambio que no puede ser previsto por ningún esquema. Según Vose [207], realmente no es tan importante saber cuántas instancias de un esquema aparecerán en la siguiente generación del algoritmo, sino saber de qué instancias se trata. El subconjunto de elementos representados por cierto esquema, normalmente, presentan una variabilidad notable del resultado de su función de ajuste y, por tanto, la siguiente generación de la población no tendrá necesariamente un resultado de su función objetivo igual al de la generación anterior.

3.5. Las técnicas de regresión

Tal y como se ha indicado anteriormente, en el contexto del *machine learning*, los métodos de regresión [208, 209] son técnicas de aprendizaje supervisado que tratan de explicar una variable numérica dependiente a partir de cierto conjunto de variables independientes. Entre las técnicas más empleadas en regresión se encuentran la regresión lineal [210], las máquinas de vectores de soporte [71] y los trazadores de regresión adaptativos multivariantes [73]. Se describen en este apartado las máquinas de vectores de soporte, dado que se trata de la metodología que, junto con

los algoritmos genéticos, constituyen la base de la metodología que se aplica en el algoritmo desarrollado en el presente proyecto de investigación.

3.5.1. Las máquinas de vectores de soporte

Las máquinas de vectores de soporte o *support vector machines* (SVM) son modelos de aprendizaje supervisados aplicables tanto a problemas de clasificación como de regresión. Dado un conjunto de datos de entrenamiento, cada uno marcado con la categoría a la que pertenece, un modelo SVM es capaz de asignar nuevos ejemplos en una categoría u otra, por lo que es un clasificador binario lineal no probabilístico. Un modelo SVM es una representación de los ejemplos como puntos en el espacio, asignados de modo que las distintas categorías se encuentran tan separadas como sea posible. Cuando se desea clasificar un nuevo elemento, se predice a qué categoría pertenece según la zona del espacio en la que se encuentre.

Además de realizar una clasificación lineal, las SVM pueden efectuar de manera eficiente clasificaciones no lineales utilizando el denominado método del núcleo [211]. Esto supone de manera implícita mapear sus entradas en espacios de características de alta dimensión. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que se puede utilizar en problemas de clasificación o regresión.

El algoritmo original de las SVM fue creado por Vapnik y Chervonenkis (1964) [212]. Años después, Boser et al. (1992) sugirieron una manera de crear clasificadores no lineales aplicando el método del núcleo a hiperplanos de margen máximo. En la actualidad, la implementación más utilizada de este método es la propuesta por Cortes y Vapnik (1995) [167].

En el caso de las SVM, cada elemento del conjunto de datos se representa como un vector. Dado un conjunto de datos en $\mathbb{R}^p$, se quiere saber si se pueden separar estos puntos con un hiperplano de dimensión $p-1$. Este hiperplano se llama clasificador lineal. Nótese que existen muchos hiperplanos que pueden clasificar los datos. Una elección razonable de hiperplano óptimo es el que representa la separación más grande, o margen, entre las dos clases. Así, se elige el hiperplano de manera que se maximiza la distancia de él hasta el punto de datos más cercano a cada uno de sus lados. Si tal hiperplano existe, se conoce como el hiperplano de margen máximo y el clasificador lineal así definido es el clasificador de margen máximo. Dado un conjunto n de puntos de entrenamiento de la forma $(\overrightarrow{x_1}, y_1), (\overrightarrow{x_2}, y_2), ..., (\overrightarrow{x_n}, y_n)$, donde los valores de y_i son 1 o -1, indicando este valor la clase a la que pertenece $\overrightarrow{x_i}$. Cada $\overrightarrow{x_i}$ es un vector real de dimensión p. Se quiere encontrar el hiperplano de margen máximo que clasifica los puntos $\overrightarrow{x_i}$ en función de los valores de y_i. Cualquier hiperplano se puede escribir como un conjunto de puntos $\overrightarrow{x}$ que satisfacen la ecuación:

$$\overrightarrow{w} \cdot \overrightarrow{x} - b = 0$$

donde w es el vector normal al hiperplano. El parámetro $\frac{b}{\|\overrightarrow{w}\|}$ determina el desplazamiento del hiperplano desde el origen a lo largo del vector normal $\overrightarrow{w}$. La Figura 1 muestra tres ejemplos de posibles hiperplanos de separación. En este ejemplo,

H_1 no separa las clases, H_2 sí las separa, pero con un pequeño margen, y H_3 las separa con el margen máximo. Si los datos de entrenamiento son linealmente separables, resulta posible seleccionar dos hiperplanos paralelos que separan las dos clases de datos, de modo que la distancia entre ellos sea lo más grande posible. La región limitada por estos dos hiperplanos se llama margen, y el hiperplano de margen máximo es el hiperplano que se encuentra a medio camino entre ellos. Estos hiperplanos se pueden describir por medio de las ecuaciones:

$$\overrightarrow{w} \cdot \overrightarrow{x} - b = 1$$
$$\overrightarrow{w} \cdot \overrightarrow{x} - b = -1$$

Geométricamente, la distancia entre estos dos hiperplanos es $\frac{2}{\|\overrightarrow{w}\|}$, de modo que maximizar esta distancia significa minimizar $\|\overrightarrow{w}\|$. Como también hay que evitar que los puntos de datos caigan en el margen, se añade la siguiente restricción,

$$\forall i \ \overrightarrow{w} \cdot \overrightarrow{x} - b \geq 1 \ si \ y_i = 1$$

O bien

$$\forall i \ \overrightarrow{w} \cdot \overrightarrow{x} - b \leq 1 \ si \ y_i = 1$$

Las ecuaciones anteriores indican que cada elemento estará situado en el lado correcto del margen. Se pueden reescribir como:

$$y_i(\overrightarrow{w} \cdot \overrightarrow{x} - b) \geq 1 \ \forall i \ 1 \leq i \leq n$$

La Figura 3.2 muestra tres ejemplos de hiperplanos que sirven pra la separación de grupos. Si bien el hiperplano H1 no consigue una separación perfecta de los círculos blancos y negros, tanto el hiperplano H2 como el H3 sí son capaces de conseguir dicha separación.

El problema de optimización asociado se puede enunciar como:

minimizar $\|\overrightarrow{w}\|$ sujeto a $y_i(\overrightarrow{w_i} \cdot \overrightarrow{x} - b) \geq 1$ para todo $i = 1, 2, ..., n$

Los vectores $\overrightarrow{w}$ y b que resuelven este problema determinan el clasificador $\overrightarrow{x} \rightarrow signo(\overrightarrow{w} \cdot \overrightarrow{x} - b)$. Una consecuencia importante y fácil de ver de esta descripción geométrica, es que el hiperplano de margen máximo queda completamente determinado por los $\overrightarrow{x_l}$ que se encuentran más cercanos a él. Estos $\overrightarrow{x_l}$ son llamados vectores de soporte.

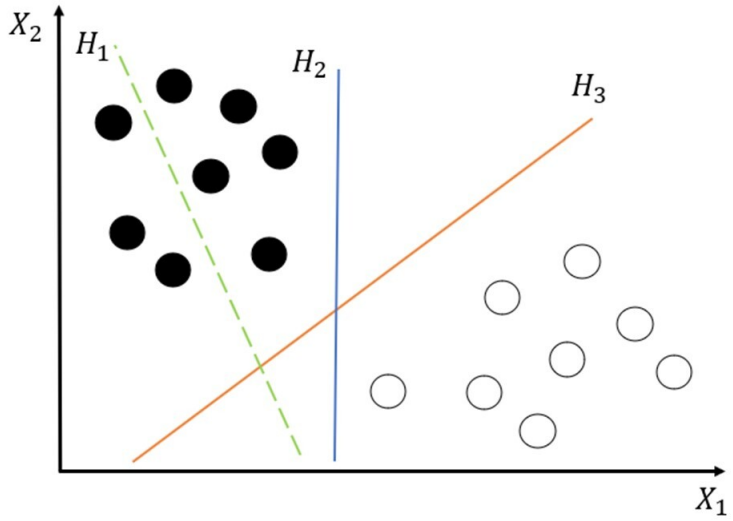


FIGURA 3.2: Ejemplos de hiperplanos para la separación de grupos.

Para extender las SVM a los casos en los que los datos no son separables linealmente, se introduce la función de pérdida de bisagra:

$$max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x} - b))$$

Esta función es cero si la restricción expuesta en la ecuación anterior se cumple. Es decir, si $\overrightarrow{x_l}$ se encuentra en el lado correcto del margen. Para los datos que se encuentran en el lado equivocado del margen, el valor de la función es proporcional a la distancia desde dicho margen. Por tanto, lo que se desea minimizar es:

$$\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - b)) + \lambda \left\| \overrightarrow{w} \right\|^2$$

donde el parámetro λ determina el equilibrio entre el aumento del tamaño del margen y asegurar que el vector $\overrightarrow{x_i}$ cae en el lado correcto de este. Por tanto, para valores suficientemente pequeños de λ, las SVM de margen blando se comportarán de forma idéntica a las de margen duro si los datos de entrada son linealmente clasificables. En la Figura 3.3 se muestra el hiperplano de margen máximo y el margen de una SVM entrenado con muestras pertenecientes a dos clases. Las muestras que se encuentran sobre el margen se llaman vectores de soporte.

El algoritmo original del hiperplano de margen máximo propuesto por Vapnik et al. en 1964 [212] construía un clasificador lineal. Sin embargo, como se señaló anteriormente, Boser et al. (1992) [70] introdujeron una manera de crear clasificadores no lineales aplicando el método del núcleo o kernel para hiperplanos de margen máximo. El algoritmo resultante es formalmente similar, excepto en que el producto escalar se sustituye por una función kernel no lineal. Esto permite al algoritmo

ajustar el hiperplano de margen máximo en un espacio característico transformado. Aunque el clasificador es un hiperplano en el espacio de características transformado, puede ser no lineal en el espacio de entrada original. Conviene señalar que, si se trabaja en un espacio de características de dimensión superior, se incrementa el error de generalización de las SVM, aunque con un número de muestras suficientes el algoritmo muestra un buen comportamiento.

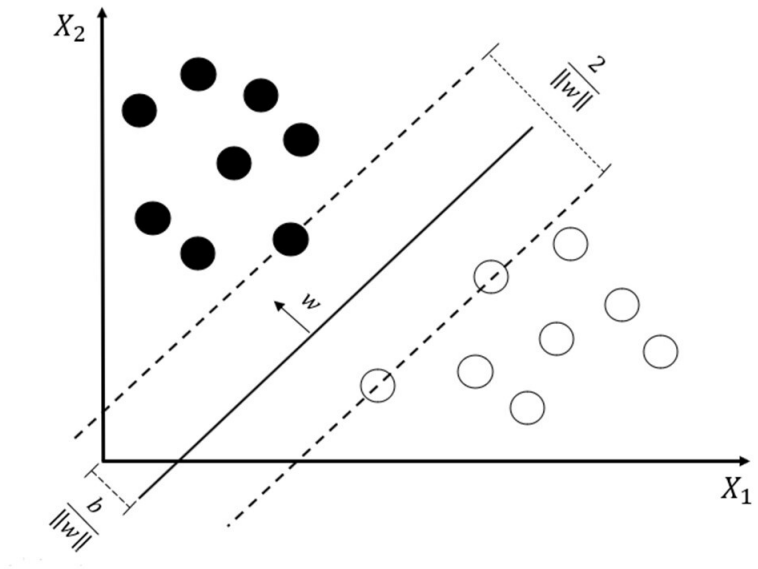


FIGURA 3.3: Hiperplano de margen máximo y margen de una SVM entrenado con muestras de dos clases.

Algunos de los núcleos más comúnmente empleados en los modelos de máquinas de vectores de soporte son los que se relacionan a continuación [213, 214]:

- Polinómico homogéneo: $k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}, \overrightarrow{x_j})^d$
- Polinómico heterogéno: $k(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i}, \overrightarrow{x_j} + d)^d$
- Función de base radial gaussiana: $k(\overrightarrow{x_i}, \overrightarrow{x_j}) = exp(-y\left|x_i - x_j\right|^2)$ con $y > 0$
- Tangente hiperbólica $k(\overrightarrow{x_i}, \overrightarrow{x_j}) = tanh((K \cdot \overrightarrow{x_i}, \overrightarrow{x_j} + c)$ para algún $K > 0$ y $c < 0$

Resulta conveniente señalar que el núcleo está relacionado con la transformación $\varphi(\overrightarrow{x_i})$ por la ecuación $k(\overrightarrow{x_i}, \overrightarrow{x_j}) = \varphi(\overrightarrow{x_i}) \cdot \varphi(\overrightarrow{x_j})$. El valor $\overrightarrow{w}$ está también en el espacio transformado, con $\overrightarrow{w_i} = \sum_i a_i \cdot y_i \cdot \varphi(\overrightarrow{x_j})$ .Los productos escalares con $\overrightarrow{w_i}$ para la clasificación se pueden calcular con el método del núcleo, es decir, $\overrightarrow{w} \cdot \varphi(\overrightarrow{x}) = \sum_i a_i \cdot y_i \cdot k(\overrightarrow{x_l}, \overrightarrow{x})$.

El cálculo del clasificador SVM de margen blando supone minimizar la expresión (James et al., 2013):

$$\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - b)) + \lambda \left\| \overrightarrow{w} \right\|^2$$

Se expone en esta ecuación el clasificador de margen blando, dado que la elección de un valor suficientemente pequeño para λ produce el clasificador de margen duro de los datos de entrada clasificables linealmente. Se detalla a continuación el enfoque clásico, que consiste en la reducción de esta ecuación a un problema de programación cuadrática.

Tal y como se introdujo en el párrafo anterior, la minimización de la expresión del clasificador de margen blando puede escribirse como un problema de optimización con restricciones con una función objetivo diferenciable en la siguiente forma: Para cada $i \in \{1, 2, ..., n\}$, se introducen las variables ξ_i y se observa que $\xi_i = max(0, 1 - y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} - b))$ si y solo si ξ_i es el número no negativo más pequeño que satisface $y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} + b) \geq 1 - \xi_i$. Por tanto, se puede escribir el problema de optimización de la siguiente forma:

$$min \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \left\| \overrightarrow{w} \right\|^2$$

sujeto a

$$y_i(\overrightarrow{w} \cdot \overrightarrow{x_i} + b) \geq 1 - \xi_i \; y \; \xi_i \geq 0$$

Este se llama problema primal. Resolviendo el dual lagrangiano del problema previo, se obtiene el problema simplificado (Wilmott, 2019):

$$max f(c_1, c_2, ...c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i (\vec{x_i} \cdot \vec{x_j}) y_i c_i$$

sujeto a

$$\sum_{i=1}^{n} c_i y_i = 0 \; y \; 0 \leq c_i \leq \frac{1}{2n\lambda} \forall i$$

Este se conoce como el problema dual. Dado que el problema de minimización dual es una función cuadrática de las c_i sujeto a restricciones lineales, se resuelve de forma eficiente usando algoritmos de programación cuadrática. También debe tenerse en cuenta que las variables c_i se definen de tal manera que:

$$\vec{w} = \sum_{i=1}^{n} c_i y_i \vec{x_i}$$

Por tanto, la eficacia de las SVM depende del núcleo elegido, de los parámetros seleccionados para el mismo y también del parámetro C de margen blando. Nótese que un valor mayor de C implica un coste mayor derivado de las muestras mal clasificadas. Un valor menor implica que las muestras mal clasificadas van a suponer un coste menor. Es decir, un valor alto de C significa que se toleran pocas muestras mal clasificadas, mientras que un valor bajo de este parámetro lo que quiere decir es que se tolera una mala clasificación de la muestra siempre que la distancia de la misma al hiperplano sea pequeña. Drucker et al. [215] propusieron una versión de las SVM para regresión. Tal y como se indicó anteriormente, el modelo producido por la clasificación basada en vectores de soporte depende solo de un subconjunto de los datos de entrenamiento, ya que la función de coste para la construcción del modelo no se preocupa por puntos de entrenamiento que se encuentran más allá del margen. Análogamente, el modelo producido por las SVM para regresión, denominadas SVR, depende solo de un subconjunto de los datos de entrenamiento, ya que la función de coste para la construcción del modelo ignora cualquier dato de entrenamiento cercano a la predicción del modelo. El entrenamiento del SVR original significa resolver:

$$min \frac{1}{2} \|\vec{w}\|^2$$

sujeto a

$$\|y_i - w_i x_i\| \leq \epsilon$$

donde x_i es una muestra de entrenamiento con el valor objetivo y_i y ϵ es un parámetro que sirve de umbral. Todas las predicciones tienen que estar dentro de un rango ϵ de las verdaderas predicciones.

A pesar de que en esta descripción de la metodología de las máquinas de vectores de soporte se ha hecho una breve introducción de estas como modelo de regresión además de como clasificador, en el caso del presente proyecto de investigación se han empleado únicamente como modelos clasificadores.

3.6. Algoritmo basado en aprendizaje automático para estudios GWAS

El algoritmo que se propone en el presente proyecto de investigación hace uso tanto de la metodología de los algoritmos genéticos como de la de las máquinas de vectores de soporte para analizar si un determinado *pathway*, que tal y como ya se ha indicado con anterioridad, en el contexto del algoritmo que aquí se desarrolla se puede considerar simplemente como un conjunto de SNPs que poseen algunas características que aconsejan su agrupación, es capaz de distinguir entre casos y controles de una cierta enfermedad o rasgo.

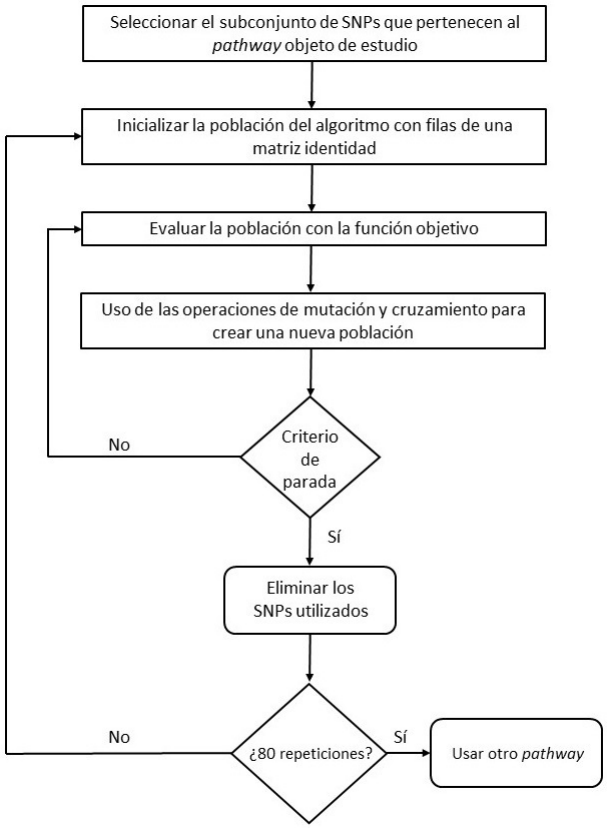


FIGURA 3.4: Flujograma del algoritmo desarrollado en el presente proyecto de investigación.

La Figura 3.4 muestra el diagrama de flujo del algoritmo propuesto. Dado que este algoritmo hace uso tanto de algoritmos genéticos como de máquinas de vectores de soporte, se le ha llamado inglés GASVeM (*genetic algorithms and support vector machines methodology*), lo que se podría traducir como metodología basada en algoritmos genéticos y máquinas de vectores de soporte.

El primer paso del algoritmo consiste en seleccionar el subconjunto de SNPs que constituyen el pathway concreto que será objeto de análisis. Esto significa que, del número total de SNPs disponibles en la base de datos (más de 300.000 en el caso de la empleada en este proyecto de investigación), la información requerida para el análisis se reduce a un subconjunto, seleccionando únicamente aquellos SNPs que pertenecen al *pathway* que es objeto de estudio.

Los miembros de la población del algoritmo genético para este análisis son, por tanto, cadenas de unos (1s) y ceros (0s) que indican qué SNPs formarán parte del modelo de máquinas de vectores de soporte que se calculará. Nótese que 1 significa que el SNP formará parte del modelo de máquinas de vectores de soporte y 0 que no lo hará. En el caso de la presente investigación, cada miembro de la población del algoritmo genético tiene la misma longitud que el número de SNPs que constituyen el *pathway* objeto de análisis. Nótese que que cada población del algoritmo genético está constituido por un conjunto de miembros, y se entrena un modelo de máquinas de vectores de soporte para cada uno de ellos.

Todos los modelos de clasificación basados en máquinas de vectores de soporte que se emplean en el presente algoritmo, se entrenan utilizando como variables de entrada los SNPs con el valor 1 y como salida, la variable rasgo que indica qué elementos son casos y cuáles son controles. Como puede verse en el diagrama de flujo, la población inicial está formada por las filas de una matriz de identidad seleccionada de forma aleatoria hasta completar el número total de individuos requeridos para la población del algoritmo genético. Así, por ejemplo, si el número de SNPs que componen el *pathway* fuera de 50, la matriz identidad que se emplearía sería de $50x50$ y si la población inicial que se emplea es de 1000 elementos, entonces se haría un muestreo con reemplazo de las filas de la matriz hasta completar una población de 1000 elementos. En caso de que la población tuviera un número de miembros menor que el de filas o columnas de la matriz, se haría igualmente empleo de la técnica de muestreo con reemplazo [216, 217].

Esto significa que en la población inicial, solo un SNP estará activo en cada uno de sus miembros. Es decir, significa que después de seleccionar como información de entrada solo aquellos SNPs que pertenecen al *pathway* objeto de análisis, la población inicial estará formada por individuos en los que solo uno de esos SNPs está activo y, posteriormente, los diferentes SNPs que pertenecen al *pathway* pueden irse manifestando en los elementos que componen las poblaciones sucesivas que se van generando. Dada la dinámica de los algoritmos genéticos, la activación y desactivación de los SNPs que forman el *pathway* viene influida por la maximización de la función objetivo. Así mismo, cabe señalar que La razón para elegir solo un SNP en cada miembro de la población inicial es que lo que se persigue es obtener los valores máximos de la función de optimización mientras se hace uso del número mínimo de SNPs y permitir que la importancia de cada SNP sea tenida en cuenta individualmente. Por tanto, para evitar que la posible influencia de un SNP enmascare la de otro, todos los elementos de la población inicial tienen activo un único SNP.

En las sucesivas poblaciones que se van generando, el número de SNPs seleccionados en cada individuo que la constituye puede ser más de uno. Tal y como se ha indicado, la evolución de la población se produce teniendo en cuenta que se persigue la maximización del valor que los individuos de la población presentan frente a la función objetivo. En el caso del algoritmo que se propone en el presente proyecto de investigación, la función objetivo consiste en calcular el área bajo la curva ROC (receiver operating characteristic) [65] que se obtiene cuando se clasifican los datos, haciendo uso de la máquina de vectores de soporte calculada para ese miembro de la población (cadena de ceros y unos), utilizando los SNP activos como variables independientes y como variables dependientes si el individuo padece o no una determinada enfermedad o bien presenta cierto rasgo, que en el caso de la base de datos empleada para la presente investigación es el cáncer colorrectal. Para evitar problemas relacionados con la epistasis, los miembros de la población que eligen más de un SNP del mismo gen tienen asignado un valor de 0 al resultado de su función objetivo.

En el párrafo anterior se ha indicado que la función objetivo es el área bajo la curva ROC. Las curvas ROC nacieron en el ámbito de la teoría de la detección de señales [69] y muy pronto su uso se extendió a otros campos como el de la epidemiología. Una curva ROC es una representación gráfica de la sensibilidad frente a la especificidad [218] para un sistema clasificador binario según se varía el umbral de discriminación [219, 220]. Así, el análisis de la curva ROC proporciona herramientas

para la selección de los mejores modelos de máquinas de vectores soporte. Es decir, permite seleccionar de manera precisa los SNPs que dan lugar a los conjuntos de datos que permiten el entrenamiento de los mejores modelos de máquinas de vectores de soporte.

Cuando se alcanza el criterio de parada, en este caso, el número máximo de ciclos permitido, se calcula el valor del área bajo la curva ROC (AUC) alcanzada y los SNPs empleados se eliminan del conjunto de SNPs disponibles. Seguidamente, el proceso comienza de nuevo buscando un nuevo subconjunto de SNPs para el que el valor del área puede ser lo más alto posible. Luego, el proceso se repite hasta que el conjunto de SNPs no empleados esté vacío o hasta que se hayan completado un total de 80 ciclos. Posteriormente, el mismo proceso se repite 1000 veces, permutando las etiquetas de casos y controles. En lo referente a estas permutaciones, cabe destacar que en este campo, parte de la literatura existente recomienda 10.000 permutaciones [221] mientras que en la literatura clásica el número de permutaciones consideradas en la mayoría de los artículos para estimar la potencia de una prueba de permutación es de 1000 [222, 223, 224], y en algunos de ellos sólo 500 [225, 226]. Finalmente, también es necesario señalar que investigaciones previas [227, 228] establecieron que 1000 es un número razonable de permutaciones para una prueba con un nivel de significación del 5 %. Finalmente, en el campo de los estudios genómicos existe un software que considera que los valores de permutación de 1000 pueden ser empleados de manera factible en GWAS [229].

Capítulo 4

Resultados y discusión

4.1. Introducción

En esta sección se detallan los resultados obtenidos de la aplicación del algoritmo desarrollado en el presente proyecto de investigación y que ya se presentó en la sección titulada «Propuesta de un algoritmo basado en aprendizaje automático para estudios de genoma amplio». Para evaluar el rendimiento del mencionado algoritmo, resulta necesario disponer de una base de datos que en el caso de la presente investigación, es la descrita en el apartado correspondiente del capítulo de Material y Metodología, así como probar con una serie de *pathways*. Concretamente, los *pathways* seleccionados fueron los que se relcionan a continuación y que ya se introdujeron en la Tabla 3.1.

- Adipocytokine signaling pathway.
- AMPK signaling pathway.
- Apelin signalling pathway.
- Colorectal cancer pathway.
- Glucagon signalling pathway.
- Enfermedad de Huntington.
- Insulin resistance.
- Insulin signalling pathway.
- Longevity regulating pathway.
- Biogénesis mitocondrial.

Entre los diez *pathways* elegidos, hay algunos de los que se conoce a través de la bibliografía la existencia de una relación muy clara con el cáncer colorectal, otros con una relación dudosa y otros sin ningún tipo de relación conocida. Esta selección no ha sido casual, sino que se hizo así con el fin de comprobar el comportamiento del algoritmo frente a todo tipo de *pathways* y las posibles asociaciones de los mismos con un rasgo.

4.2. Aplicación del diseño de experimentos al algoritmo desarrollado

Con el fin de conseguir un desempeño óptimo del algoritmo desarrollado, se ha hecho uso de la metodología de diseño de experimentos para ajustar algunos de los parámetros del mismo. Las variables que se han ajustado por medio de esta metodología han sido el número de iteraciones que empleará el algoritmo, el tamaño de la población, la probabilidad de cruzamiento y la tasa de mutación.

Antes de decidir cuáles serían los valores límite que se emplearían para cada una de las variables, se hizo un estudio previo en el que se analizó el comportamiento de cada una de las variables mencionadas en un rango mayor de variación, considerado suficiente según la experiencia del autor así como de la bibliografía relativa al estado de la técnica.

Así, en primer lugar, la Figura 4.1 muestra cómo evoluciona el área bajo la curva ROC según se incrementa el número de iteraciones entre 1000 y 10.000, con incrementos de 1000 en 1000. Cada uno de los puntos de la curva representa el valor promedio de diez repeticiones del algoritmo. Para todos estos ejemplos se han mantenido constantes el resto de variables implicadas en el algoritmo. Más concretamente, se ha hecho uso de un tamaño de la población de 5000 individuos y un valor de probabilidad de cruzamiento de 0,55, así como de una tasa de mutación de 0,01. Nótese que al realizarse este estudio previo antes del diseño de experimentos el tamaño de la población ha sido elegido teniendo en cuenta también el coste computacional del problema que nos ocupa. Además, tanto el valor elegido para la probabilidad de cruzamiento [230, 231] como para la de mutación [232, 233], han sido escogidos teniendo en cuenta estudios previos. Los resultados obtenidos se presentan en la Figura 4.1, donde se aprecia cómo a partir de las 6000 iteraciones la mejora del área bajo la curva ROC es muy pequeña.

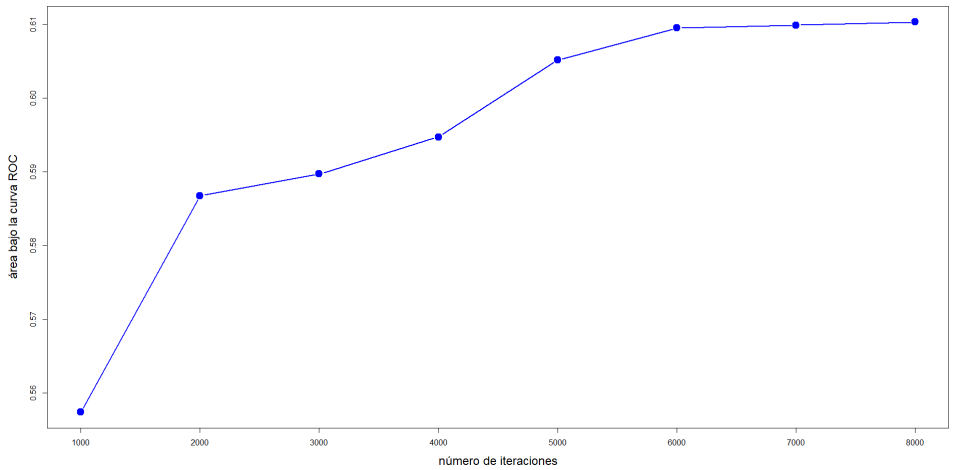


FIGURA 4.1: Área bajo la curva ROC según el número de iteraciones realizadas por el algoritmo.

Seguidamente, en la Figura 4.2 se muestra el resultado obtenido para diferentes

tamaños de población, comprendidos entre los 100 y los 15000 individuos. Todas las réplicas del algoritmo se han realizado empleando 10000 iteraciones y con valores de probabilidad de cruzamiento y mutación de 0,55 y 0,1 respectivamente, al igual que en el caso analizado con anterioridad. Así, en la mencionada figura se observa cómo la mejora en el área bajo la curva ROC es prácticamente inexistente cuando el tamaño de la población supera los 5000 individuos.

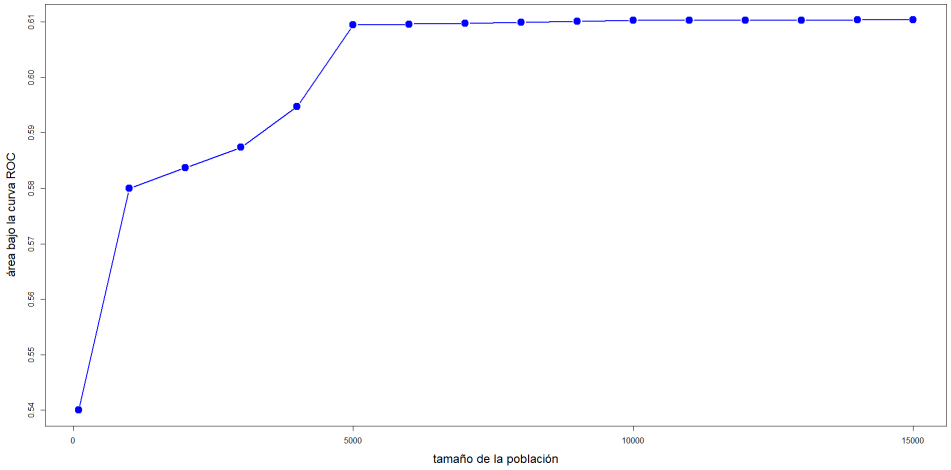


FIGURA 4.2: Área bajo la curva ROC según el tamaño de la población utilizado por el algoritmo

De forma análoga a las dos figuras vistas con anterioridad, la Figura 4.3 muestra el resultado obtenido para el valor del área bajo la curva ROC en el caso de diferentes valores de la tasa de mutación expresados en escala logarítmica. En concreto, los valores de tasa de mutación que se emplearon en el algoritmo y cuyos resultados se presentan en el gráfico son 10^{-4}, 10^{-3}, 10^{-2}, $0,1$, $0,2$ y $0,3$. Al igual que en los casos anteriores, estos ensayos, de los que se realizaron 10 repeticiones para cada uno de los niveles y se obtuvo la media, se realizaron manteniendo el resto de variables constantes. De la misma forma que en el resto de los experimentos, el tamaño de la población empleado fue de 5000 individuos, con un máximo de 10000 iteraciones en todos los casos y una probabilidad de cruzamiento del 0,55.En este caso, el valor máximo de área bajo la curva ROC se obtuvo para una tasa de mutación de 10^{-2}.

De manera equivalente a la de las tres figuras anteriores, en la Figura 4.4 se muestra la evolución de los valores del área bajo la curva ROC para distintos valores de la tasa de cruzamiento. Los valores de tasa de cruzamiento utilizados para los ensayos realizados fueron los comprendidos entre 0,1 y 1 a intervalos de 0,1. En este caso se aprecia cómo a mayor valor de la tasa de cruzamiento se consigue un mejor rendimiento del algoritmo. Es decir, un mayor valor del área bajo la curva ROC.

Teniendo en cuenta los resultados que se obtuvieron en los experimentos anteriores y que se resume en las gráficas presentadas en esta sección, se optó por efectuar un diseño de experimentos para las variables número de iteraciones, tamaño de la población, probabilidad de cruzamiento y tasa de mutación haciendo uso de los valores bajo, central y alto que se recogen en la Tabla 4.1. En el marco de este experimento, se efectuaron en orden aleatorio una serie de ensayos independientes.

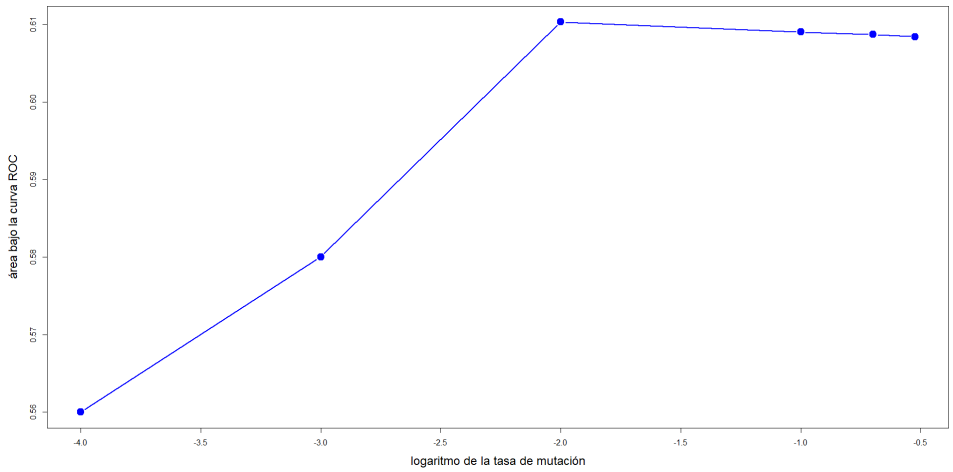


FIGURA 4.3: Área bajo la curva ROC según la tasa de mutación utilizada por el algoritmo (tasa de mutación expresada en escala logarítmica).

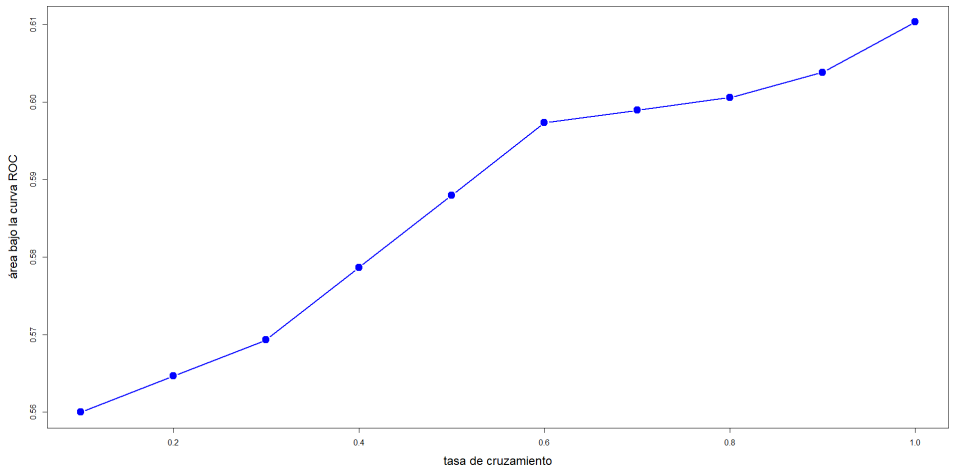


FIGURA 4.4: Área bajo la curva ROC según la tasa de cruzamiento empleada por el algoritmo.

Para cada una de las posibles combinaciones de valores que se muestran en la 4.1 se hicieron tres repeticiones del ensayo.

En primer lugar, se presentan los resultados obtenidos en los ensayos que forman parte del experimento a modo de gráficos de superfice de respuesta. Así, la Figura 4.5 presenta el gráfico de superficie de respuesta de las variables número de iteraciones y tamaño de la población frente al área bajo la curva ROC. en general, se aprecia cómo un incremento en ambas variables supone una mejor en el valor del área bajo la curva ROC. De forma análoga, en la Figura 4.6 se presenta el gráfico de superficie de respuesta de las variables número de iteraciones y tasa de mutación frente al área bajo la curva ROC. También en este caso en una primera observación los valores más altos de ambas variables dan lugar a los mayores valores de área

TABLA 4.1: Variables analizadas por medio de la metodología de diseño de experimentos y rangos de valores considerados en dichas variables.

	Bajo	Punto central	Alto
Número de iteraciones	4000	6000	8000
Tamaño de la población	1000	5000	10000
Probabilidad de cruzamiento	0,1	0,55	1
Tasa de mutación	0,001	0,01	0,1

bajo la curva ROC. La tercera y última gráfica de superficie de respuesta analizada es la que se presenta en la Figura 4.7 y en ella se presentan los valores de área bajo la curva ROC obtenidos en función de los valores de tasa de cruzamiento y tasa de mutación.

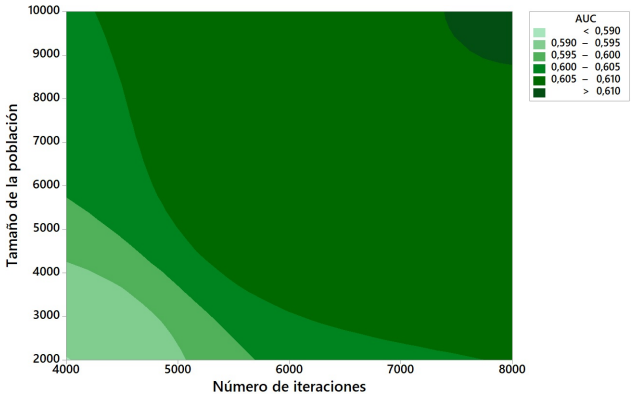


FIGURA 4.5: Gráfico de superficie de respuesta de las variables número de iteraciones y tamaño de la población frente al área bajo la curva ROC (AUC).

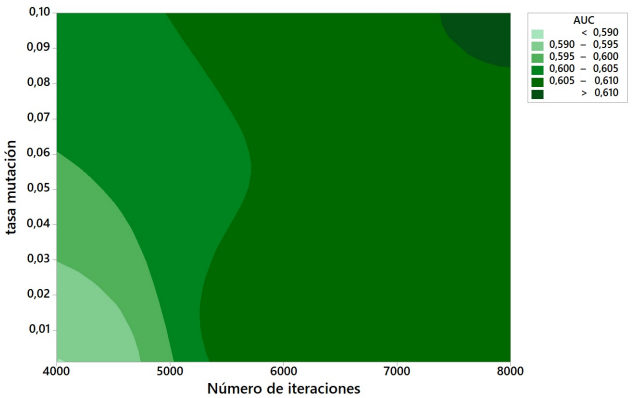


FIGURA 4.6: Gráfico de superficie de respuesta de las variables número de iteraciones y tasa de mutación frente al área bajo la curva ROC (AUC).

La Figura 4.8 muestra los gráficos de efectos principales de estas cuatro variables

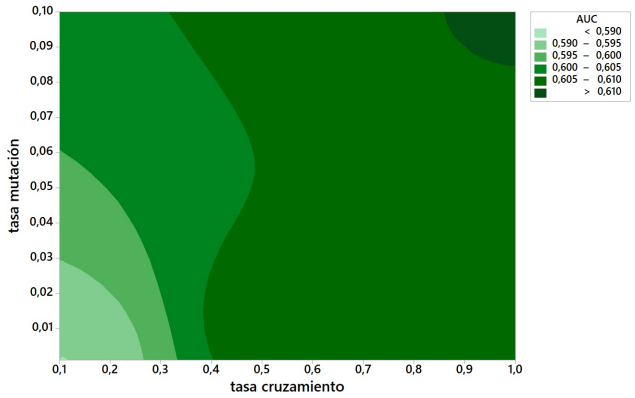


FIGURA 4.7: Gráfico de superficie de respuesta de las variables tasa de cruzamiento y tasa de mutación frente al área bajo la curva ROC (AUC).

frente a la variable dependiente que en el caso del presente problema es el área bajo la curva ROC. Según los resultados obtenidos, el número de iteraciones máximo que realizaría el algoritmo genético se fijó en 6000, dado que a partir de dicha cantidad existe únicamente un incremento porcentual muy pequeño (alrededor del 0.1 %) cuando se incrementa el número de iteraciones hasta 8000. En lo referente al tamaño de la población, se consideró que una población formada por 5500 individuos era suficiente, dado que un incremento en el número de miembros de cada población hasta, por ejemplo, 10.000 únicamente significaba una mejora por debajo del 0,2 % en el resultado del valor del área bajo la curva ROC. En lo referente a la tasa de mutación, cuyos valores se presentan en escala logarítmica, el máximo de entre los tres valores probados se alcanzó para el 1 %, punto central para esta variable del experimento. Finalmente, en el caso de la tasa de cruzamiento, se eligió un valor del 100 % dado que este era el que proporcionaba el mayor rendimiento de todos Por tanto, en vista de los resultados alcanzados en esta sección, se decidió que para todos los *pathways* analizados con el algoritmo propuesto en este proyecto de investigación, el número de iteraciones que se realizarían sería de 6000, con un tamaño de la población de 5500, una tasa de mutación del 1 % y un valor de cruzamiento del 100 %.

4.3. Aplicación del algoritmo a diferentes *pathways*

Una vez que se determinaron, en el apartado anterior, los valores óptimos que se deben de asignar a los diferentes parámetros, se aplicó el algoritmo desarrollado a todos los *pathways* que ya se indicaron en la primera sección de este capítulo. Concretamente, los parámetros que se determinaron con la ayuda del diseño de experimentos son los correspondientes al algoritmo genético. Dichos parámetros se emplearán dentro del algoritmo híbrido y son los que nuevamente se indican a continuación:

- Número de iteraciones por ciclo: 6000.
- Tamaño de la población: 5500 individuos.
- Tasa de mutación: 1 %.

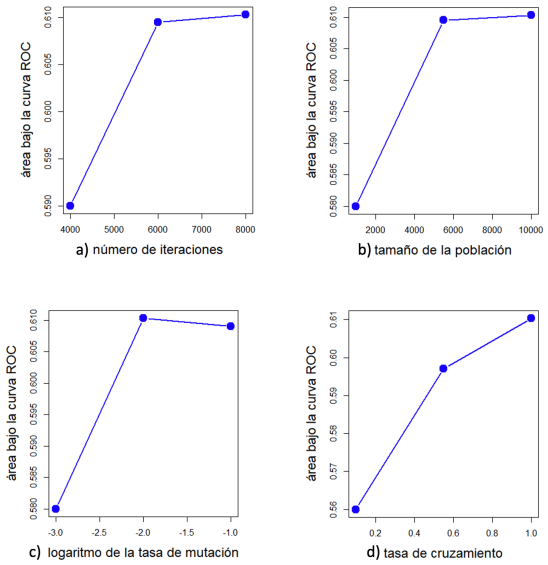


FIGURA 4.8: Gráfico de efectos principales de: (a) número de iteraciones, (b) tamaño de la población, (c) logaritmo de la tasa mutación (d) tasa de cruzamiento.

- Tasa de cruzamiento: 100 %.

En los siguientes subapartados se detalla la aplicación del algoritmo híbrido desarrollado a los *pathways* seleccionados, los cuales tienen diferentes grados de relación con el rasgo objeto de estudio, que es el cáncer colorrectal.

Tal y como ya se ha indicado en el apartado de Material y Metodología, donde se detalla el algoritmo desarrollado, en todas las aplicaciones del mismo el primer paso consiste en seleccionar como conjunto de datos inicial para la realización del estudio únicamente aquellos SNPs de la base de datos que forman parte del *pathway* objeto de estudio.

4.3.1. Aplicación del algoritmo al *adipocytokine signaling pathway*

El *pathway* denominado *adipocytokine signalling pathway* [131, 48] consta de un total de 752 SNPs presentes en la base de datos disponible para la realización de este estudio. Tal y como ya se explicó en Material y Metodología, en la población inicial, que se crea para evaluar la capacidad de clasificación del *pathway* frente al rasgo objeto de estudio, que en este caso es el cáncer colorrectal, cada uno de sus miembros tiene únicamente activo un SNP. Es decir, cada elemento de la población inicial es un vector en el que 751 de sus componentes son 0 y solo uno es 1. En este caso, al disponer de una población inicial de 5500 individuos y estar formado el pathway por 752 SNPs, se crearán las 752 cadenas en las que un único elemento está activo, y se seleccionarán 5500 individuos haciendo uso del muestreo con reemplazo sobre las 752 cadenas distintas disponibles.

Comenzado a partir de la población mencionada, los miembros de la misma se irán combinando entre sí y sufriendo procesos de mutación, dando de esta manera lugar a diversos subconjuntos de SNPs. Esta evolución de los elementos de la población se hará de forma automática, a la búsqueda de la maximización de la función

objetivo, que en el caso de este algoritmo es equivalente a la maximización del área bajo la curva ROC. Dicha evolución de la población se hace siempre teniendo en cuenta que, con el fin de evitar el fenómeno de la epistasis, no se puede incluir en un mismo miembro de la población más de un SNP perteneciente al mismo gen.

Este proceso iterativo se repite 6000 veces. Una vez que concluye el proceso indicado, se realiza la permutación de las etiquetas de casos y controles frente al rasgo objeto de estudio. El proceso de permutación se realiza 1000 veces, dando lugar a otras tantas aplicaciones del algoritmo con las etiquetas permutadas que se comparará con los resultados originales obtenidos sin dicha permutación. El propósito de realizar estas repeticiones con permutación, es comprobar si frente al rasgo objeto de estudio, los SNPs del *pathway* considerado ofrecen diferentes niveles de clasificación si se realiza una permutación de las etiquetas que los identifican como portadores del rasgo o no, con el fin de ver si existe una asociación entre el rasgo y el pathway o al menos un subconjunto de los SNPs que forman este.

La Figura 4.9 muestra los valores que se obtuvieron tras realizar 80 iteraciones, cada una de ellas de 6000 ciclos, con las etiquetas correctamente asignadas a los casos y controles frente al padecimiento o no de cáncer colorrectal, así como 5 de las 1000 ejecuciones del algoritmo que se realizaron en las mismas condiciones pero con las etiquetas de los casos y controles permutadas de manera aleatoria. Estas 5 ejecuciones que se muestran en la gráfica fueron elegidas de forma aleatoria. Nótese que al realizar las permutaciones el número total de casos y controles se mantiene invariable. Es decir, en toda permutación hay 1076 casos y 973 controles al igual que los había en el conjunto original.

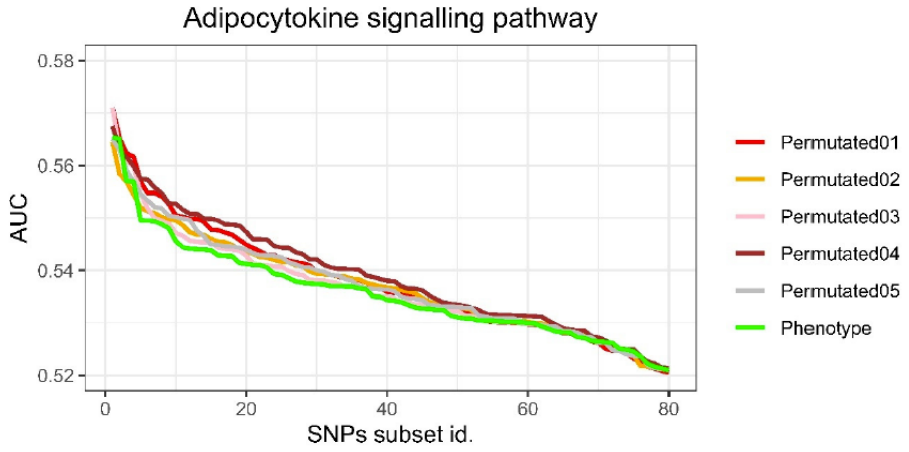


FIGURA 4.9: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *adipocytokine signaling pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

Dado que en la ejecución de cada ciclo del algoritmo se descartan los SNPs que ya han sido empleados y entran a formar parte de los modelos de máquinas de vectores de soporte otros nuevos, esto significa que no todos los SNPs fueron utilizados de forma obligatoria por el algoritmo para el proceso de clasificación de casos y controles. Así, en el caso concreto del *adipocytokine signaling pathway*, que cuenta en la

base de datos empleada en este estudio con un total de 752 SNPs, únicamente 496 fueron utilizados por todos los individuos miembros de la población en las iteraciones realizadas.

Tal y como ya se ha explicado dentro de este mismo apartado, la Figura 4.9 muestra el área bajo la curva ROC de las 80 iteraciones realizadas para el adypocitokine signalling pathway en el caso de que las etiquetas se asignan de manera correcta a los casos y controles del rasgo objeto de estudio, así como en 5 de las aplicaciones del algoritmo con las etiquetas permutadas. Estas 5 repeticiones se escogieron de manera aleatoria entre las 1000 que se efectuaron. Como era de esperar, según se van eliminando los SNPs empleados para la clasificación de casos y controles que proporcionaron mejores resultados, los que van siendo seleccionados obtienen peores resultados, siendo este el motivo por el cual según la curva se desplaza hacia la derecha disminuye la capacidad clasificadora de los modelos y por tanto el valor del área bajo la curva ROC.

Concretamente, en el caso del *adipocytokine signalling pathway*, la curva correspondiente a la ejecución en la que las etiquetas del fenotipo estaban correctamente asignadas a casos y controles sin permutación alguna, dicha curva, que está representada en verde, no parece realizar una clasificación mejor que la que se produce con el fenotipo permutado. Es decir, en la asignación de las etiquetas de casos y controles de manera aleatoria.

Finalmente, en lo referente al *pathway* objeto de análisis en este apartado, cabe destacar que los resultados de la aplicación del algoritmo al *pathway* adipocytokine signalling no muestran grandes diferencias en la capacidad de clasificación de las máquinas de vectores de soporte cuando el algoritmo aplica a la base de datos original o bien a los conjuntos de datos en los que se ha permutado la etiqueta de casos y controles. Por tanto, se trata de un *pathway* en el que no hay SNPs que permitan diferenciar a casos de controles para el rasgo objeto de análisis. Nótese que el valor medio del área bajo la curva ROC de este *pathway* en el caso sin permutar es de 0,565382.

4.3.2. Aplicación del algoritmo al *AMPK signalling pathway*

En el caso del *AMPK signaling pathway*, el valor del área bajo la curva ROC que se obtiene tras la aplicación del algoritmo a la base de datos en la que las etiquetas identifican a los individuos según su fenotipo, es ligeramente superior al valor obtenido por las bases de datos en la que se llevó a cabo una permutación de casos y controles. Además, en la mayoría de los casos, (entre el 82,5 % y el 96,5 %), el valor del área bajo la curva ROC obtenido en las iteraciones con las etiquetas correctamente asignadas según el fenotipo, son más altos que en las ejecuciones efectuadas con permutación. Por tanto, en el caso del *pathway* AMPK signaling, se ha detectado una relación débil con el cáncer colorrectal. Los resultados obtenidos para este *pathway* se muestran de manera gráfica en la Figura 4.10.

Finalmente, cabe señalar que este *pathway* está formado por un total de 1812 SNPs incluidos en la base de datos de este estudio pero, al igual que ya se señaló en el caso del *adipocytokine signaling pathway* y ocurre con todos los demás *pathways* estudiados en este proyecto de investigación, no todos ellos han sido empleados en los modelos de clasificación de máquinas de vectores de soporte. Tras el análisis

individualizado de los *pathways*, en un apartado específico de este capítulo, se proporcionará información más detallada al respecto.

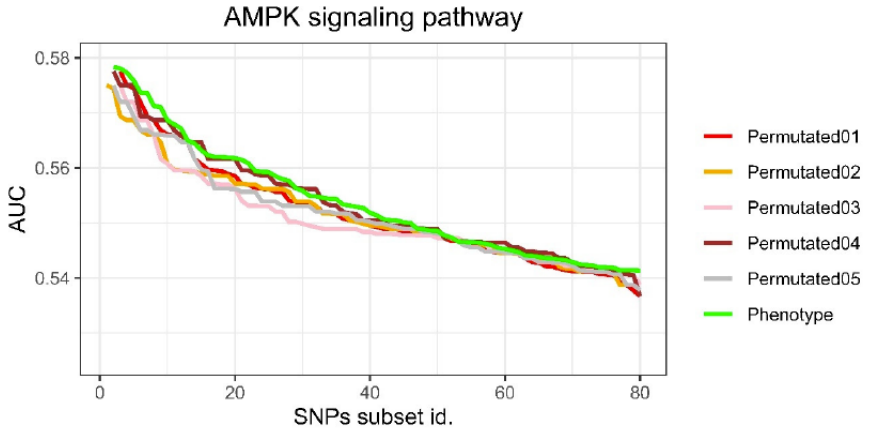


FIGURA 4.10: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *AMPK signalling pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.3. Aplicación del algoritmo al *apelin signalling pathway*

En el caso del *apelin signalling pathway*, el valor medio del área bajo la curva ROC obtenido para la ejecución del algoritmo con la base de datos original, es decir, la que hace uso de las etiquetas de casos y controles sin permutar, es claramente superior que la que se obtiene con las ejecuciones en las que los fenotipos se encuentran permutados, tal y como se puede apreciar en la Figura 4.11. Así, en este *pathway* es un 5,15 % más alta en el caso de la aplicación al fenotipo que cuando se compara con las 1000 iteraciones realizadas con las etiquetas permutadas. Por tanto, se puede decir que existe una relación clara entre el *apelin signalling pathway* y el cáncer colorrectal.

4.3.4. Aplicación del algoritmo al *pathway* asociado con el cáncer colorrectal

Como era de esperar, en el caso del *pathway* asociado específicamente con el cáncer colorrectal, el valor medio del área bajo la curva ROC es claramente superior en el caso en el que no se permutan las etiquetas si se compara con los resultados obtenidos tras realizar dicha permutación. Así, más concretamente, el valor promedio de área bajo la curva ROC es un 2.45 % más alto que en los casos en los que se permutaron las etiquetas.

En la Figura 4.18 se muestra el valor del área bajo la curva ROC obtenido en el caso sin permutación y en cinco de las 1000 repeticiones con permutación, escogidas de forma aleatoria. Esta figura confirma, desde el punto de vista gráfico, lo indicado en el párrafo anterior.

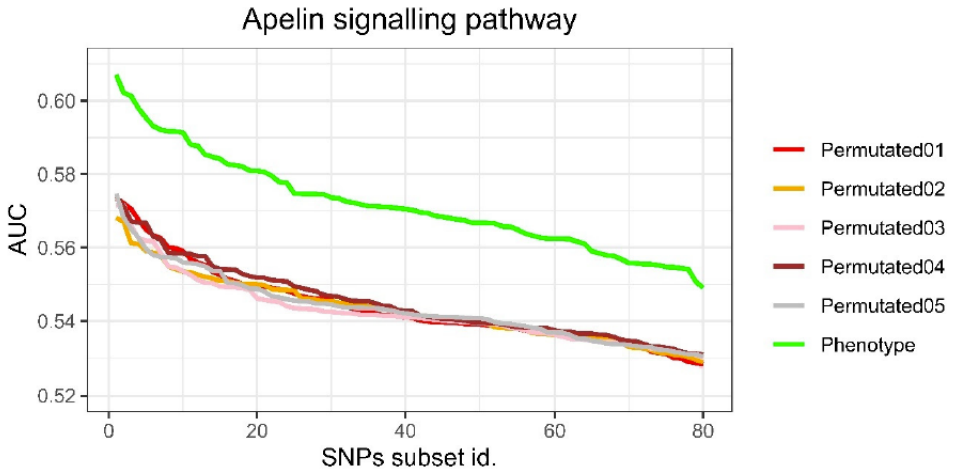


FIGURA 4.11: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *apelin signalling pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

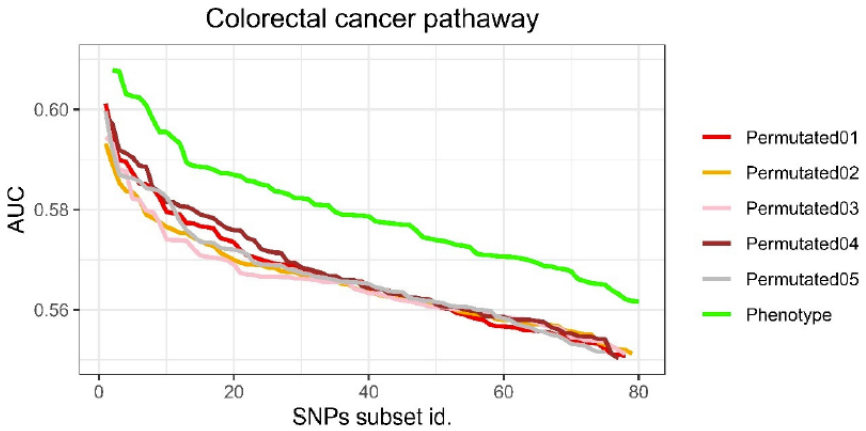


FIGURA 4.12: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *colorectal cancer pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.5. Aplicación del algoritmo al *glucagon signalling pathway*

La Figura 4.13 muestra el resultado obtenido tras la aplicación del algoritmo desarrollado al *glucagon signalling pathway*. Más concretamente, esta figura muestra los resultados que se consiguen cuando dicha aplicación se realiza teniendo en cuenta las etiquetas que identifican a casos y controles y cuando dichas etiquetas son permutadas.

A la vista de los resultados obtenidos, se puede afirmar que los SNPs pertenecientes a este *pathway* sirven para discriminar entre casos y controles frente al rasgo objeto de estudio, que es, en el caso del presente proyecto de investigación, el cáncer colorrectal.

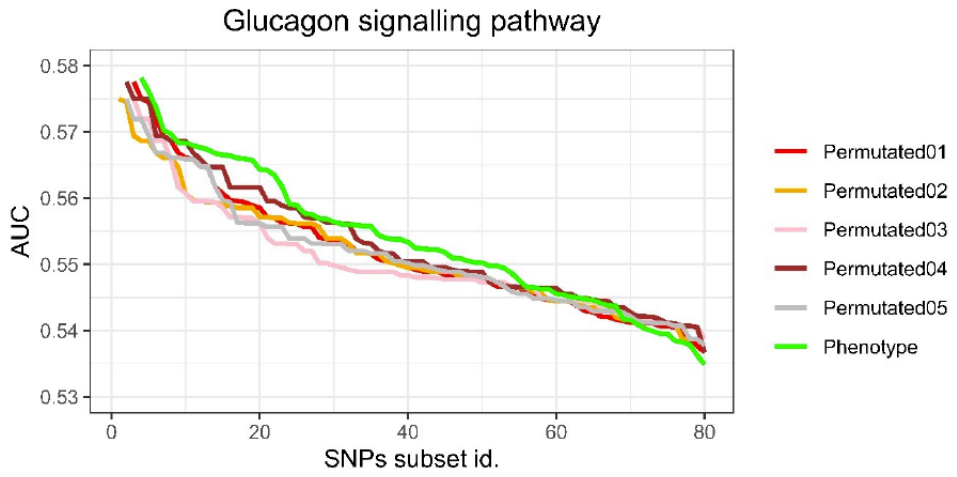


FIGURA 4.13: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *glucagon signaling pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.6. Aplicación del algoritmo al *pathway* de la enfermedad de Huntington

Los resultados obtenidos tras la aplicación del algoritmo al *pathway* relacionado con la enfermedad de Huntington se muestran en la Figura 4.14. Estos resultados presentan un comportamiento similar a los del *pathway* AMPK signalling. Así, el valor de área bajo la curva ROC obtenido en la aplicación del algoritmo a la base de datos en la que el fenotipo no ha sido permutado es ligeramente superior a los resultados que se obtienen con la permutación de fenotipo.

Por tanto, teniendo en cuenta estos resultados, se puede decir que existe una cierta relación de los SNPs de este *pathway* con el padecimiento de cáncer colorrectal, que será cuantificado más adelante dentro de este mismo capítulo de resultados.

4.3.7. Aplicación del algoritmo al *insuline resistance pathway*

En el caso del *pathway* relacionado con la resistencia a la insulina, los resultados obtenidos de la aplicación del algoritmo son los que se muestran en la Figura 4.15. Así, a la vista de estos resultados, se observa que el porcentaje de casos con permutación en los que la capacidad clasificadora es mejor que en el *pathway* con las etiquetas originales es del 29,75 %, con un valor del área bajo la curva ROC de 0,555483, para el caso de las etiquetas de casos y controles no permutados y un valor de 0,556201 en media para los casos en los que dichas etiquetas se permutaron.

Por tanto, para este *pathway* no se aprecia la existencia de relación alguna entre el mismo y el cáncer colorrectal. Por último, nótese también cómo en la Figura 4.15 tampoco parece que la curva del fenotipo esté por encima de las ejecuciones del algoritmo en las que se emplearon etiquetas permutadas y, por tanto, la representación

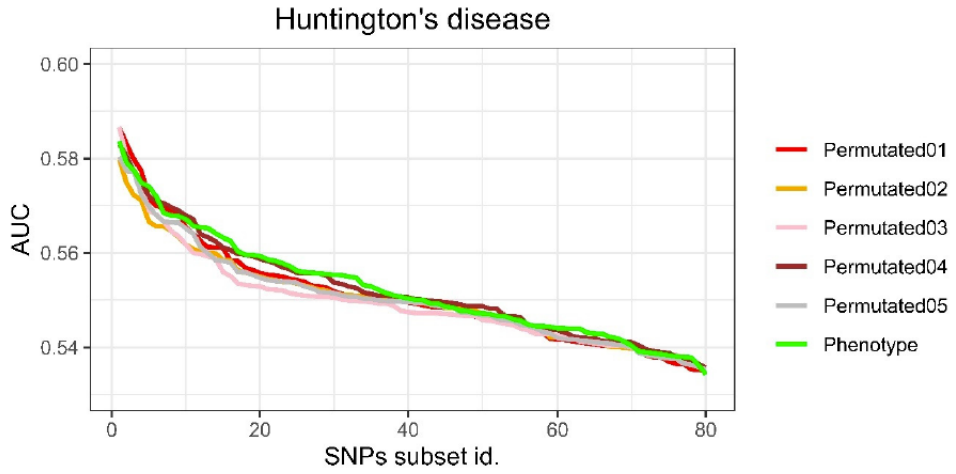


FIGURA 4.14: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *pathway* relacionado con la enfermedad de Huntington en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

gráfica confirma los resultados numéricos obtenidos.

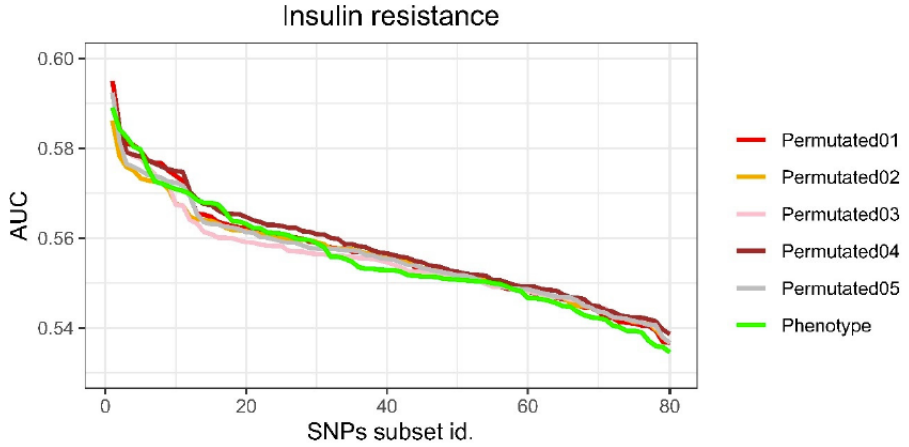


FIGURA 4.15: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *insuline resistance pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.8. Aplicación del algoritmo al *insulin signalling pathway*

La Figura 4.16 muestra los resultados obtenidos al aplicar el algoritmo desarrollado en este proyecto de investigación al *insulin signalling pathway* una vez sin aplicar permutaciones entre casos y controles y otras cinco, elegidas aleatoriamente de entre las mil realizadas, con permutación de las etiquetas de casos y controles. En este caso, los resultados obtenidos muestran una ligera capacidad clasificadora de algunos SNPs pertenecientes a este *pathway* en relación a la posible clasificación de aquellos

individuos en función de si padecen de cáncer colorrectal o no.

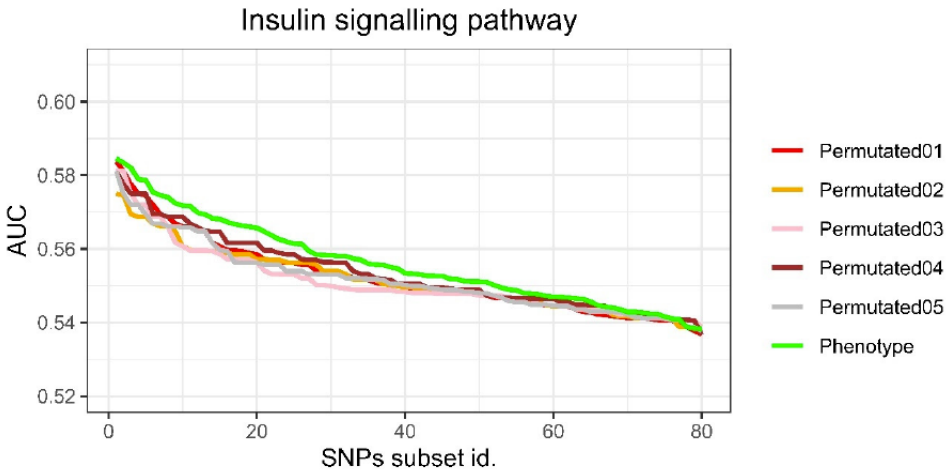


FIGURA 4.16: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *insulin signaling pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.9. Aplicación del algoritmo al *longevity regulating pathway*

En el caso del *longevity regulating pathway*, cuya curva se representa en la Figura 4.17, la situación es similar a la del *adipocytokine signaling pathway* y no se encuentra ninguna relación significativa de dicho *pathway* con el cáncer colorrectal.

Este resultado que se confirma numéricamente, también se observa de manera gráfica en la propia Figura 4.17, dado que la curva que representa la aplicación del algoritmo a los casos en los que no existe permutación del fenotipo, no se encuentra por encima de las curvas permutadas, sino que todas presentan un comportamiento muy similar.

4.3.10. Aplicación del algoritmo al *pathway* relacionado con la biogénesis mitocondrial

Los resultados obtenidos de la aplicación del algoritmo al *pathway* de la biogénesis mitocondrial se muestran en la Figura 4.18. En este caso, el valor medio del área bajo la curva ROC es claramente superior en la ejecución del algoritmo sobre la base de datos original que en las ejecuciones realizadas sobre las bases de datos que tienen los valores de las columnas del fenotipo permutadas.

Así, en el caso de este *pathway*, el valor promedio de área bajo la curva ROC supera en un 3,23 % al promedio obtenido en los casos en los que se habían permutado las etiquetas de padecimiento de cáncer colorectal en casos y controles. Nos encontramos por tanto con un *pathway* en el que algunos de sus SNPs muestran una capacidad clara de clasificación de los individuos entre aquellos que sufren cáncer

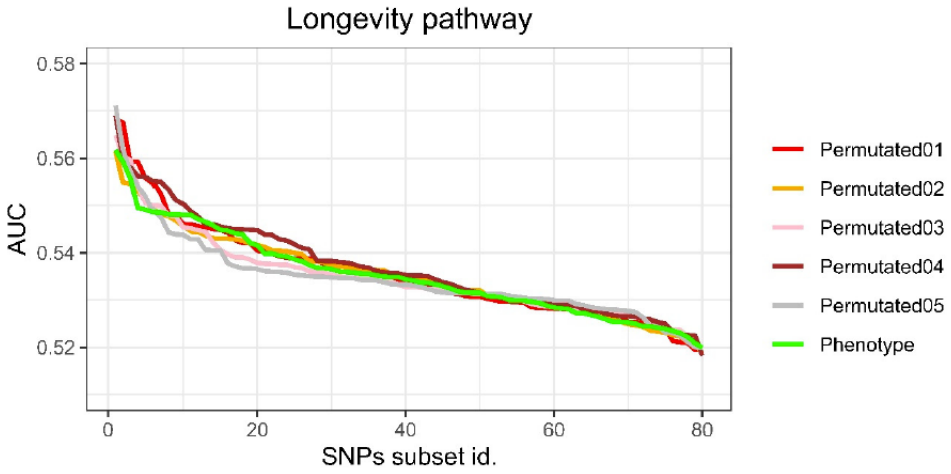


FIGURA 4.17: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *longevity regulating pathway* en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

colorrectal y los que no.

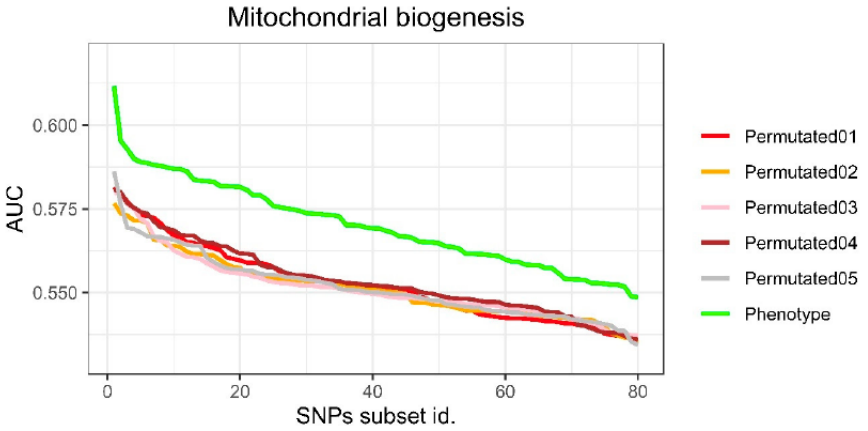


FIGURA 4.18: Valores del área bajo la curva ROC de las 80 iteraciones realizadas para el *pathway* relacionado con la biogénesis mitocondrial en el caso de los datos originales (phenotype) y en 5 permutaciones de entre las 1000 realizadas escogidas de forma aleatoria.

4.3.11. Comparación de los resultados obtenidos de la aplicación del algoritmo a los distintos *pathways* objeto de estudio

En todos los *pathways* incluidos en este estudio se ha aplicado la misma metodología. En primer lugar, para cada uno de ellos y haciendo uso de las etiquetas correspondientes a casos y controles sin permutar, se ejecutan un máximo de 6000 ciclos de un algoritmo genético en el que la función que se pretende maximizar es

el área bajo la curva ROC que se obtiene de la clasificación de casos y controles realizada por medio de máquinas de vectores de soporte. Tras la realización de todos estos ciclos, los SNPs que han sido seleccionados para el entrenamiento del modelo considerado óptimo se eliminan de la base de datos y se repite el procedimiento un total de 80 veces. Una vez terminada la ejecución de las 80 iteraciones, el proceso se repite nuevamente 1000 veces pero esta vez permutando las etiquetas que identifican a casos y controles.

Como era de esperar, en general, cuantos más SNPs significativos han sido ya eliminados, el valor del área bajo la curva ROC que se obtiene resulta más pequeño. Es decir, se produce una disminución progresiva de este índice de rendimiento. Nótese que aunque este proceso se podría repetir siempre que quedasen todavía SNPs que no hubieran sido empleados en iteraciones previas, dado que los *pathways* considerados en este estudio presentan longitudes diversas, y con el fin de conseguir por una parte una homogeneidad en los resultados obtenidos y, por otra, lograr unos tiempos de ejecución del algoritmo razonables, se decidió que el proceso finalizase después de 80 ciclos. Es decir, dado cierto *pathway* en el que se pueden haber permutado o no las etiquetas que indican el fenotipo, tras 80 ejecuciones del algoritmo, en las que en cada una de ellas se han eliminados los SNPs empleados para el entrenamiento del modelo con mejor valor de clasificación de área bajo la curva ROC, se da por finalizada la ejecución del algoritmo.

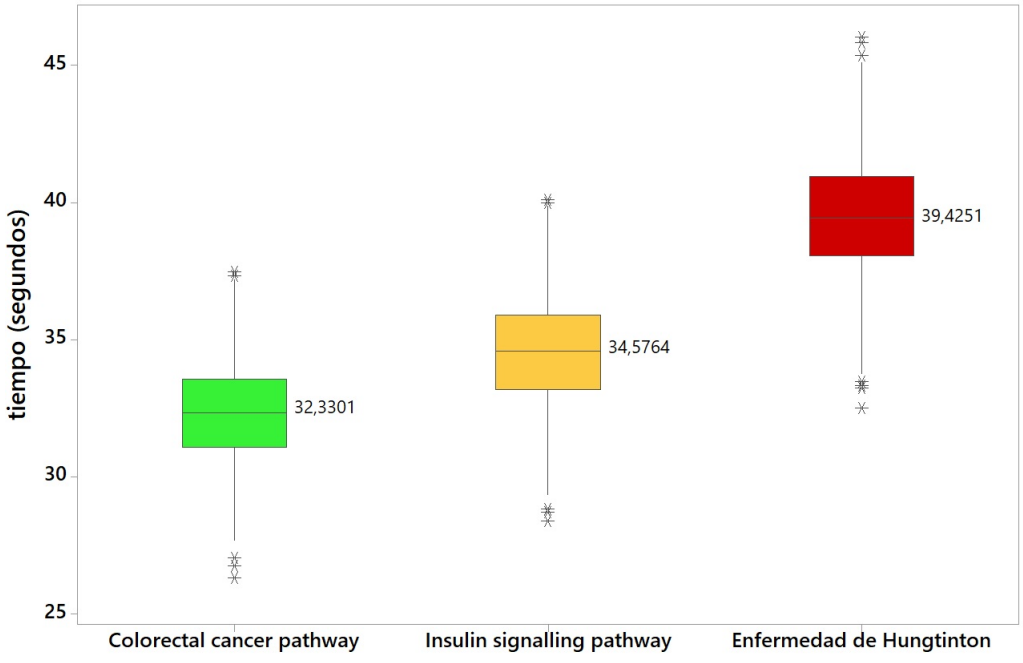


FIGURA 4.19: Diagrama de cajas de los tiempos empleados en cada una de las iteraciones efectuadas en los *pathwaty* de cáncer colorrectal, *insulin signalling pathway* y enfermedad de Hungtinton.

El tiempo medio empleado en cada una de las iteraciones realizadas fue de 34,51 segundos. Este tiempo no ha sido en promedio exactamente el mismo en todos los *pathways*, pero permite tener una idea del coste computacional del algoritmo. Así, con el fin de ofrecer una mejor aproximación del tiempo que requiere para su ejecución, la Figura 4.19 muestra el diagrama de cajas o boxplot [234, 235] de los tiempos

TABLA 4.2: Número total de SNPs que constituyen cada uno de los *pahtways* objeto de estudio. Número total de SNPs de cada uno de los *pathways* que se emplearon en alguna de las 80 iteraciones realizadas por el algoritmo en el caso en el que no se permutaron las etiquetas de casos y controles.

Nombre del pathway	**Total SNPs**	**SNPs usados**
Adipocytokine signalling pathway	752	496
AMPK signalling pathway	1812	462
Apelin signalling pathway	2525	424
Colorectal cancer pathway	813	423
Glucagon signalling pathway	1707	487
Enfermedad de Hungtinton	1980	517
Insulin resistance	1574	468
Insulin signalling pathway	1215	451
Longevity regulating pathway	1481	477
Biogénesis mitocondrial	679	438

empleados en cada una de las iteraciones efectuadas en los *pathways* de cáncer colorrectal, el *insulin signalling pathway* y el *pathway* relacionado con la enfermedad de Hungtinton. Por tanto, se ha optado por presentar en el gráfico los tiempos de estos tres pathways, dado que corresponden a los de menor tiempo en mediana (*pathway* relacionado con el cáncer colorrectal), mayor tiempo en mediana (*pathway* relacionado con la enfermedad de Hungtinton) y se ha elegido uno de los dos *pathways* con un tiempo intermedio en su mediana de ejecución, concretamente el *insulin signalling pathway*. En este diagrama, se presenta para cada uno de los tres *pathways* seleccionados el valor numérico de la mediana del tiempo de ejecución. Además, la parte superior de cada una de las cajas indica el percentil 75 del tiempo que emplearon todas las iteraciones del algoritmo, mientras que la línea inferior de la mencionada caja es el percentil 25. Nótese que por percentil 75 se entiende aquel valor que tiene por debajo al 75 % de los elementos de la población y por percentil lo mismo, pero relativo al 25 % de la misma.

Tal y como ya se ha indicado en apartados anteriores, después de ejecutar el algoritmo desarrollado para todos los *pathways* incluidos en este estudio, se ejecutó para cada uno de ellos el mismo proceso 1000 veces permutando las etiquetas de casos y controles aunque en todo caso, manteniendo el número total de 1076 casos de cáncer colorrectal y 973 controles. Los resultados numéricos más importantes obtenidos como resultado de todas estas ejecuciones del algoritmo sobre los distintos *pathways* se muestran en la Tabla 4.2 y la Tabla 4.3.

En las dos tablas indicadas en el párrafo anterior se presentan el número total de SNPs que se incluyen en cada uno de los 10 *pathways* objeto de análisis. Dado que en todos los casos el algoritmo se repitió 80 veces, en esas 80 repeticiones del caso no permutado se podría haber hecho uso de todos los SNPs disponibles. Así, la Tabla 4.2 detalla el número total de SNPs presentes en cada uno de los *pathways* estudiados junto con el número total de estos que se emplearon en las 80 iteraciones. Por ejemplo, en el caso del *adipocytokine signaling pathway*, el cual tiene un total de 752 SNPs, únicamente 496 formaron parte de los modelos considerados óptimos para la clasificación de casos y controles en alguna de las 80 iteraciones realizadas.

La Figura 4.9 muestra los valores del área bajo la curva ROC de las 80 iteraciones

TABLA 4.3: Valor promedio del área bajo la curva ROC en los distintos *pathways* objeto de estudio en el caso en el que no se realizaron permutaciones de las etiquetas de casos y controles (AUC). Valor promedio del área bajo la curva ROC de todas las iteraciones realizadas para las 1000 repeticiones del algoritmo en las que se permutaron casos y controles (AUC perm) y porcentaje de valores obtenidos del área bajo la curva ROC que superan a los valores máximos de dicho área obtenidos con el fenotipo permutado (subconjuntos ganadores).

Nombre del pathway	AUC	AUC perm	Subconjuntos ganadores
Adipocytokine signalling pathway	0,535858	0,537543	16,75 %
AMPK signalling pathway	0,564153	0,551662	89,75 %
Apelin signalling pathway	0,571761	0,543736	100 %
Colorectal cancer pathway	0,579627	0,565763	100 %
Glucagon signalling pathway	0,554759	0,552038	82,50 %
Enfermedad de Hungtinton	0,552436	0,550669	85,00 %
Insulin resistance	0,555483	0,556201	29,75 %
Insulin signalling pathway	0,556164	0,552038	96,50 %
Longevity regulating pathway	0,535285	0,53542	46,75 %
Biogénesis mitocondrial	0,570083	0,552224	100 %

realizadas para el *adipocytokine signaling pathway* en el caso en el que las etiquetas de fenotipo correspondientes a casos y controles no han sido permutadas, así como para 5 ejecuciones con permutaciones aleatorias de entre las 1000 realizadas. Para la representación gráfica y para la comparación desde el punto de visa analítico, los valores del área bajo la curva ROC se ordenan desde el más alto hasta el más bajo. En este caso, la curva denominada «de fenotipo» y que se representa en verde, no parece ser capaz de clasificar casos y controles de mejor manera que las permutaciones mostradas. Nótese también que como se recoge en la Tabla 4.3, el valor medio del área bajo la curva ROC de los 80 ciclos del fenotipo es de 0,535858, mientras que en el caso de las instancias con permutación de casos y controles su valor es de 0,537543, lo que quiere decir que es un 0.31 % inferior. La columna denominada subconjuntos ganadores indica el porcentaje de veces que el valor del área bajo la curva ROC arrojado por la curva correspondiente al fenotipo es superior al de las curvas permutadas.

En el caso del *insulin resistance pathway* ocurre algo similar, siendo el porcentaje de subconjuntos ganadores del 29,75 % y los valores del área bajo la curva ROC promedio en el caso de las ejecuciones permutadas de 0,555483 y de 0,556201 para la ejecución sin permutar (0,13 %). Por tanto, en este caso, al igual que en el *adipocytokine signaling pathway*, no parece existir relación alguna entre los dos *pathways* mencionados y el cáncer colorrectal.

De igual manera en el caso del *longevity regulating pathway*, cuya curva se representa en la Figura 4.17, la situación es similar y no se manifiesta niguna influencia significativa de este *pathway* sobre el cáncer colorrectal.

En el caso contrario se encuentran los *pathways apelin signalling*, el relacionado con la biogénesis mitocondrial y el específico del cáncer colorrectal. En estos tres casos, el valor promedio obtenido para el área bajo la curva ROC para los casos y

controles es claramente superior en el caso no permutado que en los casos calculados realizando la permutación. Así para el *apelin signalling pathway*, este es un 5,15 % superior en el caso del fenotipo cuando se compara con las soluciones obtenidas con permutación. En el caso de la biogénesis mitocondrial, está en promedio un 3,23 % por encima, y en el caso del *pathway* de cáncer colorrectal, el valor es un 2,45 % superior. Además, en todos estos casos, el valor del área bajo la curva ROC obtenido es superior en los fenotipos que en las aplicaciones realizadas con las etiquetas de casos y controles permutadas. En el caso del *apelin signaling pathway* es superior en el 100 % de los casos, lo que como ya se indicó se expresa en la tabla como «subconjuntos ganadores».

La información recogida en la tabla se complementa con las de las respectivas curvas obtenidas para el *apelin signalling pathway* (Figura 4.11), *pathway* de cáncer colorrectal (Figura 4.18) y *pathway* de biogénesis mitocondrial (Figura 4.18). En todas estas figuras se aprecia claramente cómo las curvas de los fenotipos se encuentran por encima de las permutadas.

Aunque en el caso del *AMPK signaling pathway*, la representación gráfica de la Figura 4.10 no parece mostrar de manera tan clara como en las curvas anteriores que la curva del fenotipo se encuentre por encima de las permutadas, el valor del área bajo la curva ROC es un 2,26 % más alto que el obtenido en los casos con permutación. Cabe también destacar que en el 89,75 % de los casos, los valores obtenidos son más altos en los fenotipos que en los casos permutados, lo que desde nuestro punto de vista, podría significar que existe una cierta influencia de este *pathway* sobre el rasgo objeto de estudio.

Los tres últimos *pathways* que se analizaron en el presente proyecto de investigación fueron el *glucagon signaling pathway*, cuyos resultados se muestran en la Figura 4.13, el *pathway* relacionado con la enfermedad de Huntington cuyas curvas se recogen en la Figura 4.14 y el *insulin signaling pathway* que se puede ver en la Figura 4.16. En estos tres *pathways*, de manera similar a lo que ocurre en el *AMPK signaling pathway*, el valor del área bajo la curva ROC del fenotipo es ligeramente superior que el valor promedio obtenido en los casos permutados. Además, en la mayoría de los casos, entre el 82,50 % y el 96,50 %, el área bajo la curva ROC obtenida en las iteraciones realizadas haciendo uso de los valores del fenotipo, son más altas que en el caso de las permutadas.

Finalmente,a modo de resumen, cabe indica que, teniendo en cuenta los resultados obtenidos con el algoritmo propuesto en el presente proyecto de investigación y cuyos resultados numéricos se han resumido en en las Tablas 4.2 y 4.3, se puede decir que:

- Existe una relación clara entre el cáncer colorrectal y el *pathway apelin signaling*, el *pathway* relacionado con el cáncer colorrectal y el *pathway* relacionado con la biogénesis mitocondrial.
- Existe una relación débil entre el cáncer colorrectal y el *AMPK signalling pathway*, el *glucagon signalling pathway*, el *pathway* de la enfermedad de Huntington y el *insulin signalling pathway*.

- No se ha encontrado relación alguna del cáncer colorrectal con los *pathways* *adipocytokine signaling*, *insulin resistance* y *longevity-regulating*.

4.4. Discusión

Desde el punto de vista del autor, los resultados obtenidos con los *pathways* analizados en el presente proyecto de investigación son coherentes si se comparan con los existentes en la literatura publicada hasta la fecha. Así, en lo que se refiere a los valores del área bajo la curva ROC calculados, cabe señalar que, aunque en general, y fuera del contexto de los estudios de genoma amplio podrían ser considerados como bajos, se encuentran en el rango de los valores que se suelen obtener en este tipo de estudios [236, 237, 238, 239, 240, 241].

Seguidamente, se procede a la discusión de los resultados obtenidos para cada *pathway* en comparación con los existentes en la bibliografía publicada hasta la fecha.

4.4.1. Discusión de los resultados obtenidos en relación con el *adipocytokine signalling patwhay*

La aplicación del algoritmo desarrollado al *adipocytokine signalling patwhay* nos permite afirmar que no se ha encontrado relación alguna entre este y el cáncer colorrectal. Este hallazgo es coherente con la mayoría de las publicaciones existentes hasta la fecha. Así, no parece que este *pathway* presente una asociación fuerte con el padecimiento de cáncer colorrectal, pero sí se ha encontrado relación del mismo con la mayor o menor presencia de tejido adiposo en el embarazo, la resistencia a la insulina o la hiperlipemia [131].

La literatura científica actual también pone de manifiesto que esta vía contribuye a la hipometilación del promotor y la regulación por aumento de genes de las adipocitocinas inflamatorias en los adipocitos en respuesta a la hipoxia. [242].

Entre las posibles conexiones que podrían exisir del *adipocytokine signalling patwhay* con el cáncer colorrectal, cabe destacar la influencia que tiene sobre la obesidad y diabetes. Así, se presentan alteraciones de la adipocitokina en adultos prediabéticos en los que se manifiesta la resistencia a la insulina [243, 244].

4.4.2. Discusión de los resultados obtenidos en relación con el *AMPK signalling patwhay*

En el caso del *AMPK signalling pathway*, existe un estudio en el que se encontró que el AMPK favorece la supervivencia de las células cancerígenas [245]. Más concretamente, lo que se puso de manifiesto a través de este estudio es que las células cancerígenas relacionadas con el cáncer colorrectal presentaban unos niveles más altos de genes antioxidantes y tenían unos niveles más bajos de oxígeno reactivo, que las células no relacionadas con el cáncer colorrectal. En opinión de los autores de este trabajo, este fenómeno podría ser debido a que las células de cáncer colorrectal también poseen una mayor masa mitocondrial y muestran una mayor actividad mitocondrial. En el caso del estudio de Yang et al. [245], una se observó una mayor actividad de la AMPK en estas células cancerígenas de cáncer colorrectal.

Otro estudio [246] que incluía pacientes que sufrían cáncer colorrectal en estadíos II y III, encontraron una tasa de supervivencia de los mismos a 5 años de entre el 50 % y el 87 %, poniéndose de manifiesto también en este estudio que la AMPK codificada en el gen α1 se encontraba sobreexpresada en aquellos pacientes que sufrían de cáncer colorrectal. Para estos autores, la codificación de la AMPK en el gen α1 regulaba la fosforilación de la glutatión reductasa (GSR), posiblemente a través del residuo Thr507, que potencia su actividad. De igual manera, en este trabajo, los autores también sugerían que la supresión de la expresión de la AMPK en el gen α1 por medio de vectores de nanopartículas poliméricas podría tener un efecto terapéutico beneficioso.

Finalmente, cabe indicar que en el presente proyecto de investigación el algoritmo propuesto encontró una relación considerada como débil entre el *pathway* objeto de estudio y el rasgo objeto de análisis.

4.4.3. Discusión de los resultados obtenidos en relación con el *apelin signalling patwhay*

En el caso del *apelin signalling pathway* existen algunos estudios en los que ya se encontró relevante su asociación con la aparición de cáncer colorrectal [245]. El apelin es un ligando endógeno del receptor apelin (APJ), receptor seven-transmembrane G protein-coupled receptor [245]. Dicho receptor se puede encontrar en el cerebro y también en algunos órganos periféricos como el corazón, los pulmones, los vasos sanguíneos, el tejido adiposo, etc. Una de sus funciones principales consiste en la regulación la función cardíaca y vascular, el desarrollo del corazón y en la proliferación de las células lisas del músculo vascular.

Según investigaciones conocidas, el apelin no solo está relacionado con el cáncer colorrectal, sino también con otros cánceres como el cáncer de pulmón, gastroesofágico, carcinoma hepatocelular, cáncer de próstata, cáncer de endometrio, carcinoma oral de células escamosas, cáncer de cerebro y neoangiogénesis tumoral. Esto significa que el apelin / APJ puede ser un posible objetivo terapéutico contra gran variedad de cánceres. Así por ejemplo, un estudio realizado en 2018 [247] demostró que el antagonista del receptor APJ F13A redujo significativamente la proliferación celular. Otro estudio realizado con anterioridad [248] había encontrado que el receptor del apelin se coexpresa en líneas celulares de cáncer colorrectal y su activación conduce a la inhibición de la adenilil ciclasa y la fosforilación de Akt. Para los autores de esa investigación, la apelina y su receptor podrían coexpresarse en el compartimento tumoral donde esta coexpresión subyacería a una activación constitutiva de la señalización de la apelina y crearía un bucle autocrino funcional. Según los propios autores, este fue el primer estudio que informó que el péptido del apeline se expresa altamente en adenomas y tumores de colon humano [248]. Esta coexpresión también se observó en varias líneas celulares de cáncer colorrectal. En la línea celular LoVo, los experimentos cuantitativos de reacción en cadena de la polimerasa en tiempo real (qRT-PCR) y la fosforilación de Akt inducida por el apeline confirmaron la expresión concomitante tanto del ligando como del receptor. Además, el apeline se comportó como un péptido antiapoptótico, al revertir la activación de la caspase y la degradación de la proteína poli ADP ribosa polimerasa (PARP) inducida por el inhibidor del proteasoma MG132. Otro estudio [249] que midió el apeline y su receptor mRNA, y los niveles de expresión de proteínas en tejido tumoral de 56 pacientes con adenocarcinoma colorrectal tratados quirúrgicamente y los comparó

con 27 controles sanos, encontró que los niveles séricos de receptores de apelina y sus receptores eran aumentó en los pacientes con cáncer colorrectal en comparación con los controles, lo que lleva a la conclusión de que la apelina podría ser un factor importante en la progresión del carcinoma colorrectal. El hallazgo de la vía del cáncer colorrectal como significativa por nuestro algoritmo no es sorprendente, ya que puede considerarse como la vía de referencia.

4.4.4. Discusión de los resultados obtenidos en relación con el *colorectal cancer patwhay*

El algoritmo desarrollado encontró una relación fuerte entre el rasgo objeto de estudio, que es el cáncer colorrectal y el conocido como *colorectal cancer patwhay*. Dada la naturaleza de este *pathway*, no podría ser de otra manera, existiendo en la literatura gran número de referencias bibliográficas al respecto [250, 251, 252, 253, 254].

4.4.5. Discusión de los resultados obtenidos en relación con el *glucagon signalling patwhay*

El glucagon incrementa la producción de glucosa por medio del incremento de la glicogenolisis y la gluconeogenesis en el hígado, y por medio de la reducción de la glicogénesis [255] y la glicólisis [256]. La segregación de glucagon como respuesta al consumo de alimentos, depende fundamentalmente del tipo de comida que se ingiera. Si la comida es ria en carbohidratos, los niveles de glucagon en la sangre disminuyen con el fin de prevenir un incremento del nivel circulante de glucosa. De manera contraria, en caso de que se ingiera una comida rica en proteínas, los niveles en sangre de glucagon se incrementan.

En la actualidad, es conocido que el cáncer es una de las causas de mortalidad más frecuentes entre los diabéticos [257], habiéndose realizado estudios que ponen de manifiesto la existencia de una relación entre el consumo de fármacos para el tratamiento de la diabetes y un mayor riesgo de padecimento de cáncer [258]. Estudios recientes le asignan al glucacon un papel de factor pivote implicado en la fisiopatología de la diabetes. Así, más concretamente, un estudio publicado en 2018 [259] puso de manifiesto la existencia de la expresión del receptor de glucagon en las células de cáncer colorrectal y en los tejidos afectados por cáncer colorrectal en muestras obtenidas de un grupo de pacientes. Según los resultados obtenidos en este estudio, la presencia de glucagon promueve de manera significativa el crecimiento de las células cancerígenas relacionadas con el cáncer de colon [260]. Algunos ensayos moleculares realizados mostraon que el glucagon actúa como un activador del cremiento de las célculas cancerígenas a tavés de la desactivación del AMPK y de la activación de la mitogen-activated protein kinase (MAPK) [261, 262].

Otro trabajo [263] publicado también en 2018, encontró que la existencia de una relación entre el *glucagon signalling pathway* y la presencia de cáncer de endometrio.

Por tanto, a la vista de los estudios encontrados en la literatura, resulta coherente que el algoritmo desarrollado en el presente proyecto de investigación haya encontrado una relación débil entre el cáncer colorectal y el*glucagon signalling pathway*.

4.4.6. Discusión de los resultados obtenidos en relación con el *patwhay* de la enfermedad de Huntington

La enfermedad de Huntington es una enfermedad neurodegenerativa que se considera hereditaria [264]. En sus primeras manifestaciones, sus síntomas suelen ser pequeños problemas relacionados con el estado de ánimo o con las capacidades mentales del individuo [265]. Los siguientes síntomas de la enfermedad suelen ser la falta general de coordinación así como la marcha inestable [266]. Según progresa la enfermedad, los movimientos corporales involuntarios y descoordinados se vuelven cada vez más evidentes [265]. Además, las capacidades físicas del individuo empeoran gradualmente hasta que el movimiento coordinado se vuelve difícil y la persona no puede hablar [265]. Además, en la mayoría de las ocasiones, los individuos que sufren la enfermedad acaban padeciendo demencia [266], aunque existen variaciones en los síntomas sufridos de un paciente a otro.

Los síntomas de esta enfermedad comienzan generalmente entre los 30 y los 50 años de edad, pero pueden comenzar a cualquier edad [266]. En aquellas familias en las que existen personas que la padecen, es común que la enfermedad se desarrolle a una edad más cercana en sus descendientes. Así, aproximadamente el 8 % de los casos comienzan antes de los 20 años y se conocen como enfermedad de Huntington juvenil, que generalmente se presenta con síntomas similares a los de la enfermedad de Parkinson [266, 267].

En el análisis de la bibliografía existente, no se ha encontrado estudio alguno que relacione la enfermedad de Huntington con el mayor o menor padecimiento de cáncer colorectal, lo que resulta coherente con los resultados obtenidos de la aplicación del algoritmo en el presente proyecto de investigación, en el que este *pathway* no mostró relación alguna con el rasgo objeto de estudio.

Otro estudio publicado en 2002 [268] puso de manifiesto que la enfermedad de Huntington proporciona pistas sobre el padecimiento de cáncer y podría ser un buen marcador para ciertos tipos de cáncer como el colorrectal. Así, resulta posible afirmar que, en general, aquellas personas que sufren de la enfermedad de Huntington tienen una menor propension a padecer cáncer [269]. Aunque ha sido posible encontrar un estudio centrado en el análisis de la relación existente entre la enfermedad de Huntington y el cáncer de próstata [270], hasta el momento no se conoce estudio de esta naturaleza en el que se haya establecido algún tipo de asociación, ya bien sea positiva o negativa, con el padecimiento de cáncer colorrectal.

4.4.7. Discusión de los resultados obtenidos en relación con el *patwhay* de resistencia a la insulina

La resistencia a la insulina es un fenómeno que se produce cuando las células de los músculos, la grasa y el hígado no responden bien a la insulina y no pueden utilizar la glucosa de la sangre para obtener energía [271, 272]. Para compensarlo, el páncreas debe producir más insulina [273]. Con el tiempo, se produce un aumento del nivel de azúcar en sangre. La resistencia a la insulina se caracteriza por una serie de problemas tales como la obesidad [274], la presión arterial alta, el colesterol alto y la diabetes tipo 2 [275]. Se trata de un síndrome con una alta prevalencia en población adulta [276, 277]. En los análisis realizados en el presente proyecto de investigación, a través del algoritmo desarrollado, no se ha encontrado relación alguna

entre este *pathway* y el rasgo objeto de estudio que es el cáncer colorrectal.

Una de las características fundamentales de la resistencia a la insulina es que se trata de una enfermedad silenciosa, sin síntomas en sus primeras etapas y que solo puede detectarse a través de un análisis de sangre [278, 279].

Existen una serie de factores de riesgo y causas de la resistencia a la insulina. Algunas de las más comunes son las que se relacionan a continuación [280, 281]:

- Presencia de grasa abdominal y obesidad.
- Estilo de vida poco activo.
- Dieta rica en carbohidratos.
- Diabetes gestacional.
- Hígado graso no alcohólico.
- Antecedentes familiares de diabetes.
- Síndrome del ovario poliquístico.
- Consumo de tabaco.
- Edad superior a 45 años.
- Trastornos hormonales como el síndrome de Cushing y la acromegalia.
- Medicamentos tales como los esteroides, antipsicóticos y medicamentos contra el VIH.
- Apnea del sueño.

4.4.8. Discusión de los resultados obtenidos en relación con el *insulin signalling patwhay*

El *insulin signaling pathway* es otro de los *pathways* en los que ha sido posible encontrar en la bibliografía una asociación moderada con el cáncer colorrectal. Así, existen estudios en los que se han presentado evidencias de que la modificación de los niveles de insulina debidos a la dieta podrían afectar al padecimiento de cáncer colorrectal [282]. Resultados similares a este se encontraron también en otro estudio realizado con una muestra de mujeres posmenopáusicas [283]. Aunque es posible encontrar muchos estudios con resultados en esta línea [284, 285], y es conocido que las variantes en los *pathways* metabólicos pueden interactuar con factores relacionados con los hábitos dietéticos tales como los ácidos grasos y presentar influencia en el riesgo de cáncer colorrectal [286, 287, 288], es necesario señalar que este tipo mecanismo de interacción no se comprende por completo en la actualidad [289].

4.4.9. Discusión de los resultados obtenidos en relación con el *longevity regulating pathway*

En el presente proyecto de investigación, haciendo uso del algoritmo desarrollado en el mismo, no se ha encontrado relación alguna entre el *longevity regulating pathway* y el padecimiento de cáncer colorrectal. Esto está en consonancia con los resultados obtenidos hasta el momento en la literatura [290, 152], dado que si bien el mencionado pathway se sabe que modula procesos como la autofagia, la síntesis de proteínas, la detección de nutrientes, la función mitocondrial o el estrés oxidativo, hasta el momento no existen hallazgos relativos a su relación con el cáncer colorrectal.

4.4.10. Discusión de los resultados obtenidos en el *patwhay* relacionado con la biogénesis mitocondrial

Las mitocondrias son orgánulos semiautónomos que participan en la metabolización de la energía, la producción de radicales libres y en la apoptosis [291]. Además de los núcleos, las mitocondrías son los únicos orgánulos celulares que tienen su propio genoma y un mecanismo genético propio [292, 293]. La biogénesis mitocondrial es un proceso esencial a través del cual se obtiene una nueva mitocondria, y es uno de los procesos que requieren de coordinación entre los genomas nucleares y mitocondriales [294]. La mitocondria, al igual que la mayoría de los procesos relacionados con la misma, están íntimamente ligados con la génesis del cáncer [295]. Es por este motivo por el que no debe resultar extraño que en el presente proyecto de investigación se haya encontrado, con la ayuda del algoritmo propuesto, una relación fuerte entre este *pathway* y el cáncer colorrectal.

Por tanto, para el conocimiento de la génesis del cáncer resulta esencial el estudio de la biogénesis mitocondrial, así como conocer cómo se comportan estos orgánulos durante el proceso tumoral. Así, es sabido que la progresión del cáncer colorrectal se encuentra íntimamente ligada con la alteración mitocondrial, el incremento de la producción de radicales libres mitocondriales y el estrés oxidativo [296].

Entre las aportaciones en esta línea de investigación, merece la pena destacar un artículo de revisión bibliográfica [297], el cual encontró en investigaciones previas que una expresión alterada del PGC1α1 modifica los riesgos de padecimiento de cáncer colorrectal y de la biogénesis mitocondrial, la cual se regula con la ayuda del PGC1α1. Según los resultados puestos de manifiesto en los artículos analizados en ese estudio, es coherente que el algoritmo propuesto en este proyecto de investigación haya encontrado una relación entre el *pathway* de la biogénesis mitocondrial y el cáncer colorrectal.

Capítulo 5

Conclusiones

El presente proyecto de investigación introduce un algoritmo completamente nuevo para el análisis de *pathways*. Dicho algoritmo ha sido denominado GASVeM, acrónimo que en inglés quiere decir *genetic algorithms support vector machines methodology*. Este nombre se puede traducir al español como metodología basada en algoritmos genéticos y máquinas de vectores de soporte. Tal y como su nombre indica, este algoritmo está basado en la combinación de la metodología de los algoritmos genéticos con las máquinas de vectores de sporte. Tanto la explicación del algoritmo como los resultados de su aplicación a una serie de *pathways* se publicaron en el año 2021 en la revista Mathematics, perteneciente al primer cuartil del campo de las matemáticas aplicadas del índice Journal Citation Reports. Más concretamente, dicha publicación que se recoge a modo de anexo de la presente memoria fue la que se detalla a continuación [298]:

Díez Díaz, F., Sánchez Lasheras, F., Moreno, V., Moratalla-Navarro, F., Molina de la Torre, A. J., and Martín Sánchez, V. (2021). GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. Mathematics, 9(6), 654.

Si bien la metodología propuesta permite conocer exactamente qué SNPs de cada uno de los *pathways* se utilizan para la clasificación de casos y controles según el rasgo objeto de estudio, en el análisis de los resultados no se ha llegado al análisis de cada uno de los SNPs de manera individual, sino que se ha estudiado únicamente si a partir del uso de un subconjunto de SNPs pertenecientes a cierto *pathway* resulta posible hacer una clasificación del conjunto de individuos objeto de estudio en función del rasgo analizado.

Por otra parte, tambien resulta conveniente señalar que aunque los resultados obtenidos parecen prometedores, como ocurre normalmente a la hora de aplicar metodologías basadas en *machine learning* en el contexto de los estudios de genoma amplio, resulta complejo encontrar una explicación biológica de los resultados obtenidos. A pesar de esto y a partir de la revisión de la literatura existente, ha sido posible para el autor encontrar en estudios previos publicados en revistas internacionales de prestigio hallazgos que parecen confirmar los resultados arrojados por el nuevo algoritmo desarrollado en el marco de este proyecto de investigación.

En opinión del autor, dada la falta de un metodología de aprendizaje automático que pueda considerarse como el *gold standard* en el contexto de los estudios de genoma amplio, el método que aquí se propone se considera que podría ser de gran interés para futuras investigaciones en el campo de los estudios de genoma amplio. En esta línea, y basándonos en los resultados obtenidos, podría ser de interés por

una parte, trabajar en el estudio de las capacidades de clasificación en aquellos casos en los que se combine la información genética perteneciente a varios *pathways* y, por otra parte, trabajar con el fin de poder aplicar esta metodología a otro conjunto de pacientes que presenten alguna otra enfermedad o rasgo que sean de interés.

Desde mi punto de vista, al igual que otros autores, también considero que la aplicación de las metodologías de *machine learning* en el campo de los estudios de genoma amplio se encuentra en su infancia [79]. Por tanto, todavía nos encontramos lejos de fijar algún tipo de metodología basada en aprendizaje automático que pueda considerarse como un *gold standard* y que permita producir resultados que sean fácilmente contrastables e interpretables desde el punto de vista biológico.

Además, resulta necesario señalar que, dadas las características propias de las bases de datos empleadas en este tipo de estudios, donde se dispone habitualmente de un número muy elevado de SNPs (columnas) en comparación con casos y controles (filas), nos encontramos ante un tipo de problema que es difícil de tratar desde el punto de vista de las técnicas de *machine learning* y esto hace que, en el caso de la metodología que se ha presentado en este proyecto de investigación, se realice una selección previa de los SNPs que se considerarán teniendo en cuenta el *pathway* objeto de análisis. En nuestra opinión, este problema puede tener un impacto elevado en la reproducibilidad de los resultados cuando se aplique un mismo algoritmo a dos bases de datos diferentes.

Por tanto, en virtud de lo expuesto en los párrafos anteriores, considero que es posible afirmar que la hipótesis de este trabajo de investigación, en la que se afirma que es posible el desarrollo de una metodología basada en técnicas de *machine learning* capaz de detectar qué SNPs son relevantes para la manifestación de cierto rasgo en el contexto de los estudios de genoma amplio, ha sido validada.

Finalmente, cabe destacar que somos conscientes de que la traslación de los resultados obtenidos con este método a la práctica clínica con el fin de poder personalizar tratamientos y diagnósticos requiere todavía de mayores avances en esta línea de investigación. Sin embargo, consideramos que el método ha demostrado, como era el objetivo de este proyecto de investigación, una buena capacidad para discriminar entre los *pathways* que están asociados con el rasgo y los que no, mediante la elección de un conjunto limitado de SNPs.

5.1. Conclusiones específicas

Se relacionan a continuación las principales conclusiones específicas obtenidas del presente proyecto de investigación. A través de estas conclusiones específicas, se considera que se ha alcanzado tanto el objetivo general del presente proyecto de investigación como los objetivos específicos:

- Los resultados obtenidos demuestran que es posible la construcción de nuevos algoritmos de análisis de *pathways* basado en técnicas de aprendizaje automático que consideren relaciones multivariadas entre todos los SNPs.
- En este proyecto de investigación se ha presentado una nueva metodología basada en técnicas de *machine learning* que no solo sirven para la selección de

los SNPs más relevantes dentro de cierto *pathway* previamente definido para la detección del cáncer colorrectal, sino que se trata de una metodología que, por una parte, también podría emplearse como un método preliminar para la reducción dimensional del *pathway* analizado. Desde el punto de vista del autor, esta aplicación también podría ser de interés en un futuro.

- Tal y como se ha venido indicando a lo largo de todo el presente proyecto de investigación, aunque el algoritmo desarrollado se ha aplicado a una base de datos con casos y controles provenientes de un estudio de cáncer colorrectal, dicho algoritmo sería de aplicación a cualquier otro tipo de cáncer o rasgo que fuera de interés.

- Aunque desde un punto de vista biológico, en la actualidad resulta bastante difícil encontrar una relación directa entre parte de los SNPs seleccionados por los algoritmos propuestos con el cáncer, en nuestro entendimiento es de interés la colaboración de equipos interdisciplinares que puedan abordar este tipo de problemas desde diferentes puntos de vista, fundamentalmente desde la genética y aprendizaje automático.

- Así pues, aunque los algoritmos presentados superan a métodos anteriores con los que se comparan, también tiene algunas limitaciones, fundamentalmente relacionadas con la consideración del fenómeno de la epítasis que no ha sido tenido en cuenta, pues aumentaba de forma exponencial la complejidad del problema. Actualmente, los autores continúan desarrollando algoritmos híbridos que mejorarían los resultados de los algoritmos existentes de aplicación a los estudios de genoma amplio.

Capítulo 6

Líneas futuras de investigación

6.1. Introducción

En este proyecto de investigación se ha presentado y validado una nueva metodología que no solo sirven para la selección de los SNPs más relevantes dentro de cierto *pathway* previamente definido para la detección del rasgo objeto de interés, en este caso el cáncer colorrectal, sino que también podría emplearse como un método preliminar para la reducción dimensional del *pathway* analizado. Desde el punto de vista del autor, esta aplicación también podría ser de interés en un futuro.

Seguidamente se exponen algunas posibles líneas de investigación que se podrían desarrollar en los próximos años tomando como base lo expuesto en el presente proyecto de investigación.

6.1.1. Reducción dimensional

Tal y como ya se ha indicado en la sección correspondiente a Resultados y Discusión, no todos los SNPs que forman cada *pathway* han sido utilizados por los modelos de clasificación óptimos. Por tanto, partiendo del algoritmo desarrollado en el presente proyecto de investigación sería posible trabajar en métodos que permitan realizar una reducción dimensional de los *pathways*.

6.1.2. Análisis conjunto de *pathways* e interacción entre los mismos

El método desarrollado se podría modificar con el fin de poder considerar de manera conjunta más de un *pathway* y, además, también sería posible modificar el algoritmo de forma que sea capaz de tener en cuenta las posibles interacciones entre *pathways*.

6.1.3. Aplicación de metodologías basadas en *deep learning*

Desde el punto de vista del doctorando, el *deep learning* o aprendizaje profundo [299, 300] es una herramienta de aprendizaje automático de gran interés para su aplicación en estudios de genoma amplio. En concreto, se propone como línea futura de investigación la aplicación de algoritmos basados en redes neuronales convolucionales a los estudios de genoma amplio.

Con el fin de sacar provecho a las metodologías de aprendizaje profundo, resultará necesario investigar también en sistemas de computación de alto rendimiento o computación en paralelo, dado que la ejecución de los sistemas de *deep learning* necesitan de esta tecnología para poder obtener resultados en unos tiempos razonables.

6.1.4. Medicina personalizada

En la actualidad se entiende por medicina personalizada la práctica de la medicina que hace uso del perfil genético de un individuo para guiar las decisiones tomadas en relación con la prevención, diagnóstico y tratamiento de las enfermedades. Dado que el conocimiento del perfil genético de un paciente puede ayudar a los médicos a seleccionar la medicina o la terapia más adecuada para un paciente así como para administrar la dosis o el régimen adecuados, en nuestra opinión resulta de gran interés investigar cómo se puede relacionar, a través de algoritmos de machine learning, la información genómica disponible con la reacción al tratamiento de un paciente.

Bibliografía

[1] *Genome-Wide Association Studies*. Springer Singapore, 2019. DOI: 10.1007/978-981-13-8177-5. URL: http://dx.doi.org/10.1007/978-981-13-8177-5.

[2] T. H. MORGAN. «RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE». En: *Science* 34.873 (sep. de 1911), 384–384. DOI: 10.1126/science.34.873.384. URL: http://dx.doi.org/10.1126/science.34.873.384.

[3] *National Human Genome Research Institute. The Human Genome Project*. https://www.genome.gov/human-genome-project. Accessed: 2021-06-12.

[4] Victor Lyamichev y col. «Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes». En: *Nature Biotechnology* 17.3 (mar. de 1999), 292–296. DOI: 10.1038/7044. URL: http://dx.doi.org/10.1038/7044.

[5] Jacklyn N. Hellwege y col. «Population Stratification in Genetic Association Studies». En: *Current Protocols in Human Genetics* 95.1 (oct. de 2017). DOI: 10.1002/cphg.48. URL: http://dx.doi.org/10.1002/cphg.48.

[6] Lon R Cardon y Lyle J Palmer. «Population stratification and spurious allelic association». En: *The Lancet* 361.9357 (feb. de 2003), 598–604. DOI: 10.1016/s0140-6736(03)12520-2. URL: http://dx.doi.org/10.1016/S0140-6736(03)12520-2.

[7] B. Devlin y Kathryn Roeder. «Genomic Control for Association Studies». En: *Biometrics* 55.4 (dic. de 1999), 997–1004. DOI: 10.1111/j.0006-341x.1999.00997.x. URL: http://dx.doi.org/10.1111/j.0006-341X.1999.00997.x.

[8] P. Armitage. «Tests for Linear Trends in Proportions and Frequencies». En: *Biometrics* 11.3 (sep. de 1955), pág. 375. DOI: 10.2307/3001775. URL: http://dx.doi.org/10.2307/3001775.

[9] William G. Cochran. «Some Methods for Strengthening the Common 2 Tests». En: *Biometrics* 10.4 (dic. de 1954), pág. 417. DOI: 10.2307/3001616. URL: http://dx.doi.org/10.2307/3001616.

[10] Bao-Zhu Yang y col. «Practical population group assignment with selected informative markers: Characteristics and properties of Bayesian clustering via STRUCTURE». En: *Genetic Epidemiology* 28.4 (2005), 302–312. DOI: 10.1002/gepi.20070. URL: http://dx.doi.org/10.1002/gepi.20070.

[11] Alkes L Price y col. «Principal components analysis corrects for stratification in genome-wide association studies». En: *Nature Genetics* 38.8 (jul. de 2006), 904–909. DOI: 10.1038/ng1847. URL: http://dx.doi.org/10.1038/ng1847.

[12] Leroy Hood y Lee Rowen. «The human genome project: big science transforms biology and medicine». En: *Genome Medicine* 5.9 (2013), pág. 79. DOI: 10.1186/gm483. URL: http://dx.doi.org/10.1186/gm483.

[13] The International HapMap Consortium. «A haplotype map of the human genome». En: *Nature* 437.7063 (oct. de 2005), págs. 1299-1320. DOI: 10.1038/nature04226. URL: http://dx.doi.org/10.1038/nature04226.

[14] Kouichi Ozaki y col. «Functional SNPs in the lymphotoxin- gene that are associated with susceptibility to myocardial infarction». En: *Nature Genetics* 32.4 (nov. de 2002), 650–654. DOI: 10.1038/ng1047. URL: http://dx.doi.org/10.1038/ng1047.

[15] Montgomery Slatkin. «Linkage disequilibrium — understanding the evolutionary past and mapping the medical future». En: *Nature Reviews Genetics* 9.6 (jun. de 2008), 477–485. DOI: 10.1038/nrg2361. URL: http://dx.doi.org/10.1038/nrg2361.

[16] Laura Fachal y Alison M Dunning. «From candidate gene studies to GWAS and post-GWAS analyses in breast cancer». En: *Current Opinion in Genetics Development* 30 (feb. de 2015), 32–41. DOI: 10.1016/j.gde.2015.01.004. URL: http://dx.doi.org/10.1016/j.gde.2015.01.004.

[17] Sarah E. Bergen y Tracey L. Petryshen. «Genome-wide association studies of schizophrenia». En: *Current Opinion in Psychiatry* 25.2 (mar. de 2012), 76–82. DOI: 10.1097/yco.0b013e32835035dd. URL: http://dx.doi.org/10.1097/YCO.0b013e32835035dd.

[18] «A second generation human haplotype map of over 3.1 million SNPs». En: *Nature* 449.7164 (oct. de 2007), 851–861. DOI: 10.1038/nature06258. URL: http://dx.doi.org/10.1038/nature06258.

[19] Francis Robert y Jerry Pelletier. «Exploring the Impact of Single-Nucleotide Polymorphisms on Translation». En: *Frontiers in Genetics* 9 (oct. de 2018). DOI: 10.3389/fgene.2018.00507. URL: http://dx.doi.org/10.3389/fgene.2018.00507.

[20] H. Haga y col. «Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome». En: *Journal of Human Genetics* 47.11 (nov. de 2002), 0605–0610. DOI: 10.1007/s100380200092. URL: http://dx.doi.org/10.1007/s100380200092.

[21] A. DeWan y col. «HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration». En: *Science* 314.5801 (nov. de 2006), 989–992. DOI: 10.1126/science.1133807. URL: http://dx.doi.org/10.1126/science.1133807.

[22] Daniel C. Koboldt y col. «The Next-Generation Sequencing Revolution and Its Impact on Genomics». En: *Cell* 155.1 (sep. de 2013), 27–38. DOI: 10.1016/j.cell.2013.09.006. URL: http://dx.doi.org/10.1016/j.cell.2013.09.006.

[23] Andreas Ziegler y col. «Introduction to genetic analysis workshop 17 summaries». En: *Genetic Epidemiology* 35.S1 (2011), S1–S4. DOI: 10.1002/gepi.20641. URL: http://dx.doi.org/10.1002/gepi.20641.

[24] Holly K. Tabor, Neil J. Risch y Richard M. Myers. «Candidate-gene approaches for studying complex genetic traits: practical considerations». En: *Nature Reviews Genetics* 3.5 (mayo de 2002), 391–397. DOI: 10.1038/nrg796. URL: http://dx.doi.org/10.1038/nrg796.

[25] Christoph Lippert y col. «An Exhaustive Epistatic SNP Association Analysis on Expanded Wellcome Trust Data». En: *Scientific Reports* 3.1 (ene. de 2013). DOI: 10.1038/srep01099. URL: http://dx.doi.org/10.1038/srep01099.

[26] Chao Ning y col. «Efficient multivariate analysis algorithms for longitudinal genome-wide association studies». En: *Bioinformatics* 35.23 (mayo de 2019). Ed. por John Hancock, 4879–4885. DOI: 10.1093/bioinformatics/btz304. URL: http://dx.doi.org/10.1093/bioinformatics/btz304.

[27] Hai Lin y col. «RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants». En: *Genome Biology* 20.1 (nov. de 2019). DOI: 10.1186/s13059-019-1847-4. URL: http://dx.doi.org/10.1186/s13059-019-1847-4.

[28] Alberto Romagnoni y col. «Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data». En: *Scientific Reports* 9.1 (jul. de 2019). DOI: 10.1038/s41598-019-46649-z. URL: http://dx.doi.org/10.1038/s41598-019-46649-z.

[29] Eun Pyo Hong y Ji Wan Park. «Sample Size and Statistical Power Calculation in Genetic Association Studies». En: *Genomics Informatics* 10.2 (2012), pág. 117. DOI: 10.5808/gi.2012.10.2.117. URL: http://dx.doi.org/10.5808/GI.2012.10.2.117.

[30] Karl Pearson. «X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling». En: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (jul. de 1900), págs. 157-175. DOI: 10.1080/14786440009463897. URL: http://dx.doi.org/10.1080/14786440009463897.

[31] Chao-Ying Joanne Peng, Kuk Lida Lee y Gary M. Ingersoll. «An Introduction to Logistic Regression Analysis and Reporting». En: *The Journal of Educational Research* 96.1 (sep. de 2002), págs. 3-14. DOI: 10.1080/00220670209598786. URL: http://dx.doi.org/10.1080/00220670209598786.

[32] Joseph O Ogutu, Torben Schulz-Streeck y Hans-Peter Piepho. «Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions». En: *BMC Proceedings* 6.S2 (mayo de 2012). DOI: 10.1186/1753-6561-6-s2-s10. URL: http://dx.doi.org/10.1186/1753-6561-6-S2-S10.

[33] M. B. WILK y R. GNANADESIKAN. «Probability plotting methods for the analysis for the analysis of data». En: *Biometrika* 55.1 (1968), págs. 1-17. DOI: 10.1093/biomet/55.1.1. URL: http://dx.doi.org/10.1093/biomet/55.1.1.

[34] Jian Yang y col. «Genomic inflation factors under polygenic inheritance». En: *European Journal of Human Genetics* 19.7 (mar. de 2011), págs. 807-812. DOI: 10.1038/ejhg.2011.39. URL: http://dx.doi.org/10.1038/ejhg.2011.39.

[35] Olive Jean Dunn. «Multiple Comparisons among Means». En: *Journal of the American Statistical Association* 56.293 (mar. de 1961), págs. 52-64. DOI: 10.1080/01621459.1961.10482090. URL: http://dx.doi.org/10.1080/01621459.1961.10482090.

[36] Pak C. Sham y Shaun M. Purcell. «Statistical power and significance testing in large-scale genetic studies». En: *Nature Reviews Genetics* 15.5 (abr. de 2014), 335–346. DOI: 10.1038/nrg3706. URL: http://dx.doi.org/10.1038/nrg3706.

[37] Xiaoyi Gao y col. «Avoiding the high Bonferroni penalty in genome-wide association studies». En: *Genetic Epidemiology* (2009), n/a-n/a. DOI: 10.1002/gepi.20430. URL: http://dx.doi.org/10.1002/gepi.20430.

[38] Teri A. Manolio y col. «Finding the missing heritability of complex diseases». En: *Nature* 461.7265 (oct. de 2009), 747–753. DOI: 10.1038/nature08494. URL: http://dx.doi.org/10.1038/nature08494.

[39] S. Lee, M. C. Wu y X. Lin. «Optimal tests for rare variant effects in sequencing association studies». En: *Biostatistics* 13.4 (jun. de 2012), 762–775. DOI: 10.1093/biostatistics/kxs014. URL: http://dx.doi.org/10.1093/biostatistics/kxs014.

[40] Jason H. Moore. «The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases». En: *Human Heredity* 56.1-3 (2003), págs. 73-82. DOI: 10.1159/000073735. URL: http://dx.doi.org/10.1159/000073735.

[41] Carlos de Céspedes Montealegre. «Epistasis y persistencia de la enfermedad». En: *Acta Médica Costarricense* 61.4 (jul. de 2020). DOI: 10.51481/amc.v61i4.1043. URL: http://dx.doi.org/10.51481/amc.v61i4.1043.

[42] H. J. Cordell. «Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans». En: *Human Molecular Genetics* 11.20 (oct. de 2002), 2463–2468. DOI: 10.1093/hmg/11.20.2463. URL: http://dx.doi.org/10.1093/hmg/11.20.2463.

[43] S. J. Russell y P. Norvig. *Artificial Intelligence: A Modern Approach (3.ª ed.)* Upper Saddle River, Nueva Jersey: Prentice Hall, 2010.

[44] W. S. McCulloch y W. Pitts. «A logical calculus of the ideas immanent in nervous activity». En: *Bulletin of Mathematical Biophysics* 5 (1943), págs. 115-133.

[45] A. Turing. «Computing Machinery and Intelligence». En: *Mind, LIX* (1950), págs. 433-460.

[46] Arthur L. Samuel. «Some Studies in Machine Learning Using the Game of Checkers». En: *IBM Journal of Research and Development* 44 (1959), págs. 206-226.

[47] F. Rosenblatt. «The perceptron: A probabilistic model for information storage and organization in the brain». En: *Psychological Review* 65 (1958), págs. 386-408.

[48] B. Widrow y M. A. Lehr. «30 years of adaptive neural networks: perceptron, madaline, and backpropagation». En: *Proceedings of the IEEE*. 78, 1990, págs. 1415-1442.

[49] Galileo Galilei. «Il Saggiatore (en italiano». En: *The Controversy on the Comets of 1618* D. O'Malley (1623).

[50] G. Boole. *The Mathematical Analysis of Logic, Being an Essay towards a Calculus of Deductive Reasoning*. Londres: Macmillan, Barclay Macmillan, 1847.

[51] G. Frege. *Begriffsschrift*. Alemania, Halle: Verlag von Louis Nebert, 1879.

[52] A. Tarski. «Truth and Proof». En: *Scientific American* 220 (1969), págs. 63-77.

[53] Kurt Gödel. «Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I». En: *Monatshefte für Mathematik und Physik* 38-38.1 (dic. de 1931), págs. 173-198. DOI: 10.1007/bf01700692. URL: http://dx.doi.org/10.1007/BF01700692.

[54] S. C. Kleene. *Introducción a Metamathematics*. Amsterdam: North-Holland, 1952.

[55] P. E. Ceruzzi. *A History of Modern Computers*. Cambridge: MIT Press, 1998.

[56] M. R. Garey y D. S. Johnson. *Computers and Intractability – A Guide to the Theory of NP-Completeness*. Nueva York: W. H. Freeman y Company, 1979.

[57] *Gerolamo Cardano*. https://www.biografiasyvidas.com/biografia/c/cardano.htm. Accessed: 2021-6-12.

[58] A Hald. *History of probability and statistics and their applications before 1750: Hald/history of probability & statistics*. en. Nashville, TN, Estados Unidos de América: John Wiley & Sons, 1990.

[59] T Bayes. «An Essay towards solving a Problem in the Doctrine of Chances». En: *Philosophical Transactions of the Royal Society of London* 53 (1763), págs. 370-418.

[60] P Wilmott. *Machine learning: An applied mathematics introduction*. Reino Unido: Panda Ohana Publishing, 2019.

[61] J. B. MacQueen. «Some Methods for classification and Analysis of Multivariate Observations». En: *En Le Cam*. Ed. por L. M. y J. Neyman. Proceedings of 5th Berkeley Symposium on Mathematical Statistics y Probability . Berkeley, California: University of California Press, 1967, págs. 281-297.

[62] S. P. Lloyd. «Least squares quantization in PCM». En: *IEEE Transactions on Information Theory* 28 (1982), págs. 129-137.

[63] J. Holland. *Adaptation in Natural and Artificial Systems*. Cambridge, Massachusetts: MIT Press, 1975.

[64] D. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Boston: Addison-Wesley, 1989.

[65] T. Hastie, R. Tibshirani y J. Friedman. *The Elements of Statistical Learning (2.ª ed.)* Nueva York: Springer Verlag, 2008.

[66] L. Rokach y O. Maimon. «Top-down induction of decision trees classifiers-a survey». En: *IEEE Transactions on Systems, Man, and Cybernetics, Part C* 35 (2005), págs. 476-487.

[67] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington D. C.: Spartan Books, 1961.

[68] T. Kohonen. «Self-Organized Formation of Topologically Correct Feature Maps». En: *Biological Cybernetics* 43 (1982), págs. 59-69.

[69] P. McCullagh y J. A. Nelder. *An outline of generalized linear models. (2.ª ed.)* Boca Ratón, Florida: Chapman Hall, 1999.

[70] B. E. Boser, I. M. Guyon y V. N. Vapnik. «A Training Algorithm for Optimal Margin Classifiers». En: *En COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (1992), págs. 144-152.

[71] C. Cortes y V. N. Vapnik. «Support-vector networks». En: *Machine Learning* 20 (1995), págs. 273-297.

[72] A. Ben-Hur y col. «Support vector clustering». En: *Journal of Machine Learning Research* 2 (2001), págs. 125-137.

[73] J. H. Friedman. «Multivariate Adaptive Regression Splines». En: *The Annals of Statistics* 19 (1991), págs. 1-67.

[74] Silke Szymczak y col. «Machine learning in genome-wide association studies». En: *Genetic Epidemiology* 33.S1 (2009), S51–S57. DOI: 10.1002/gepi.20473. URL: http://dx.doi.org/10.1002/gepi.20473.

[75] Bettina Mieth y col. «Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies». En: *Scientific Reports* 6.1 (nov. de 2016). DOI: 10.1038/srep36671. URL: http://dx.doi.org/10.1038/srep36671.

[76] Sangseob Leem y col. «Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure». En: *Computational Biology and Chemistry* 50 (jun. de 2014), 19–28. DOI: 10.1016/j.compbiolchem.2014.01.005. URL: http://dx.doi.org/10.1016/j.compbiolchem.2014.01.005.

[77] Guillaume Paré, Shihong Mao y Wei Q. Deng. «A machine-learning heuristic to improve gene score prediction of polygenic traits». En: *Scientific Reports* 7.1 (oct. de 2017). DOI: 10.1038/s41598-017-13056-1. URL: http://dx.doi.org/10.1038/s41598-017-13056-1.

[78] Dimitrios Vitsios y Slavé Petrovski. *Stochastic semi-supervised learning to prioritise genes from high-throughput genomic screens*. Mayo de 2019. DOI: 10.1101/655449. URL: http://dx.doi.org/10.1101/655449.

[79] Hannah L. Nicholls y col. «Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci». En: *Frontiers in Genetics* 11 (abr. de 2020). DOI: 10.3389/fgene.2020.00350. URL: http://dx.doi.org/10.3389/fgene.2020.00350.

[80] Tony Burdett y col. *GWAS Catalog*. https://www.ebi.ac.uk/gwas/. Accessed: 2021-6-13.

[81] Juan Pablo Lewinger y col. «Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation». En: *Genetic Epidemiology* 31.8 (dic. de 2007), 871–882. DOI: 10.1002/gepi.20248. URL: http://dx.doi.org/10.1002/gepi.20248.

[82] Giorgos Mountrakis, Jungho Im y Caesar Ogole. «Support vector machines in remote sensing: A review». En: *ISPRS Journal of Photogrammetry and Remote Sensing* 66.3 (mayo de 2011), págs. 247-259. DOI: 10.1016/j.isprsjprs.2010.11.001. URL: http://dx.doi.org/10.1016/j.isprsjprs.2010.11.001.

[83] Leo Breiman. En: *Machine Learning* 45.1 (2001), págs. 5-32. DOI: 10.1023/a:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[84] Alexey Natekin y Alois Knoll. «Gradient boosting machines, a tutorial». En: *Frontiers in Neurorobotics* 7 (2013). DOI: 10.3389/fnbot.2013.00021. URL: http://dx.doi.org/10.3389/fnbot.2013.00021.

[85] Adrian Iustin Georgevici y Marius Terblanche. «Neural networks and deep learning: a brief introduction». En: *Intensive Care Medicine* 45.5 (feb. de 2019), págs. 712-714. DOI: 10.1007/s00134-019-05537-w. URL: http://dx.doi.org/10.1007/s00134-019-05537-w.

[86] Brooke L. Fridley y col. «A Latent Model for Prioritization of SNPs for Functional Studies». En: *PLoS ONE* 6.6 (jun. de 2011). Ed. por Stein Aerts, e20764. DOI: 10.1371/journal.pone.0020764. URL: http://dx.doi.org/10.1371/journal.pone.0020764.

[87] Somayeh Kafaie, Yuanzhu Chen y Ting Hu. «A network approach to prioritizing susceptibility genes for genome-wide association studies». En: *Genetic Epidemiology* 43.5 (mar. de 2019), págs. 477-491. DOI: 10.1002/gepi.22198. URL: http://dx.doi.org/10.1002/gepi.22198.

[88] Rahul C Deo y col. «Prioritizing causal disease genes using unbiased genomic features». En: *Genome Biology* 15.12 (dic. de 2014). DOI: 10.1186/s13059-014-0534-8. URL: http://dx.doi.org/10.1186/s13059-014-0534-8.

[89] Usman Roshan y col. «Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest». En: *Nucleic Acids Research* 39.9 (feb. de 2011), e62-e62. DOI: 10.1093/nar/gkr064. URL: http://dx.doi.org/10.1093/nar/gkr064.

[90] Ivan Merelli y col. «SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS». En: *BMC Bioinformatics* 14.S1 (ene. de 2013). DOI: 10.1186/1471-2105-14-s1-s9. URL: http://dx.doi.org/10.1186/1471-2105-14-S1-S9.

[91] Max Schubach y col. «Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants». En: *Scientific Reports* 7.1 (jun. de 2017). DOI: 10.1038/s41598-017-03011-5. URL: http://dx.doi.org/10.1038/s41598-017-03011-5.

[92] Atlas Khan, Qian Liu y Kai Wang. «iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes». En: *BMC Bioinformatics* 19.S17 (dic. de 2018). DOI: 10.1186/s12859-018-2469-7. URL: http://dx.doi.org/10.1186/s12859-018-2469-7.

[93] Jian Zhou y col. «Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk». En: *Nature Genetics* 50.8 (jul. de 2018), págs. 1171-1179. DOI: 10.1038/s41588-018-0160-6. URL: http://dx.doi.org/10.1038/s41588-018-0160-6.

[94] Ozgur Demir-Kavuk y col. «Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features». En: *BMC Bioinformatics* 12.1 (2011), pág. 412. DOI: 10.1186/1471-2105-12-412. URL: http://dx.doi.org/10.1186/1471-2105-12-412.

[95] Ofer Isakov, Iris Dotan y Shay Ben-Shachar. «Machine Learning–Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease». En: *Inflammatory Bowel Diseases* 23.9 (sep. de 2017), págs. 1516-1523. DOI: 10.1097/mib.0000000000001222. URL: http://dx.doi.org/10.1097/MIB.0000000000001222.

[96] François R. Jornayvaz y Gerald I. Shulman. «Regulation of mitochondrial biogenesis». En: *Essays in Biochemistry* 47 (jun. de 2010). Ed. por Guy C. Brown y Michael P. Murphy, págs. 69-84. DOI: 10.1042/bse0470069. URL: http://dx.doi.org/10.1042/bse0470069.

[97] Kyle Gettler y col. «Prioritizing Crohn's disease genes by integrating association signals with gene expression implicates monocyte subsets». En: *Genes Immunity* 20.7 (ene. de 2019), págs. 577-588. DOI: 10.1038/s41435-019-0059-y. URL: http://dx.doi.org/10.1038/s41435-019-0059-y.

[98] J. A. Nelder y R. W. M. Wedderburn. «Generalized Linear Models». En: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pág. 370. DOI: 10.2307/2344614. URL: http://dx.doi.org/10.2307/2344614.

[99] Hui Zou y Trevor Hastie. «Regularization and variable selection via the elastic net». En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (abr. de 2005), págs. 301-320. DOI: 10.1111/j.1467-9868.2005.00503.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

[100] Enrique Medina. «Enfermedad inflamatoria intestinal (I): clasificación, etiología y clínica». En: *Anales de Pediatría Continuada* 11.2 (mar. de 2013), págs. 59-67. DOI: 10.1016/s1696-2818(13)70120-3. URL: http://dx.doi.org/10.1016/S1696-2818(13)70120-3.

[101] Lei Chen. «Curse of Dimensionality». En: *Encyclopedia of Database Systems*. Springer US, 2009, págs. 545-546. DOI: 10.1007/978-0-387-39940-9_133. URL: http://dx.doi.org/10.1007/978-0-387-39940-9_133.

[102] *Gene Ontology Resource*. http://geneontology.org/. Accessed: 2021-6-13.

[103] *Home - IMPC*. en. https://www.mousephenotype.org/. Accessed: 2021-6-13. Nov. de 2018.

[104] *Human Phenotype Ontology*. https://hpo.jax.org/app/. Accessed: 2021-6-13.

[105] *OMIM - Online Mendelian Inheritance in Man*. https://www.omim.org/. Accessed: 2021-6-13.

[106] Thomas G. Dietterich. «Ensemble Methods in Machine Learning». En: *Multiple Classifier Systems*. Springer Berlin Heidelberg, 2000, págs. 1-15. DOI: 10.1007/3-540-45014-9_1. URL: http://dx.doi.org/10.1007/3-540-45014-9_1.

[107] Malgorzata Maciukiewicz y col. «GWAS-based machine learning approach to predict duloxetine response in major depressive disorder». En: *Journal of Psychiatric Research* 99 (abr. de 2018), págs. 62-68. DOI: 10.1016/j.jpsychires.2017.12.009. URL: http://dx.doi.org/10.1016/j.jpsychires.2017.12.009.

[108] Alex J. Smola y Bernhard Schölkopf. «A tutorial on support vector regression». En: *Statistics and Computing* 14.3 (ago. de 2004), págs. 199-222. DOI: 10.1023/b:stco.0000035301.49549.88. URL: http://dx.doi.org/10.1023/b:stco.0000035301.49549.88.

[109] Maryam M Najafabadi y col. «Deep learning applications and challenges in big data analytics». En: *Journal of Big Data* 2.1 (feb. de 2015). DOI: 10.1186/s40537-014-0007-7. URL: http://dx.doi.org/10.1186/s40537-014-0007-7.

[110] Nay Aung y col. «Genome-Wide Analysis of Left Ventricular Image-Derived Phenotypes Identifies Fourteen Loci Associated With Cardiac Morphogenesis and Heart Failure Development». En: *Circulation* 140.16 (oct. de 2019), págs. 1318-1330. DOI: 10.1161/circulationaha.119.041161. URL: http://dx.doi.org/10.1161/CIRCULATIONAHA.119.041161.

[111] Nils Hampe y col. «Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey». En: *Frontiers in Cardiovascular Medicine* 6 (nov. de 2019). DOI: 10.3389/fcvm.2019.00172. URL: http://dx.doi.org/10.3389/fcvm.2019.00172.

[112] Wanwen Zeng, Mengmeng Wu y Rui Jiang. «Prediction of enhancer-promoter interactions via natural language processing». En: *BMC Genomics* 19.S2 (mayo de 2018). DOI: 10.1186/s12864-018-4459-6. URL: http://dx.doi.org/10.1186/s12864-018-4459-6.

[113] John W. Davey y col. «Genome-wide genetic marker discovery and genotyping using next-generation sequencing». En: *Nature Reviews Genetics* 12.7 (jun. de 2011), págs. 499-510. DOI: 10.1038/nrg3012. URL: http://dx.doi.org/10.1038/nrg3012.

[114] Jian-Bing Fan, Mark S. Chee y Kevin L. Gunderson. «Highly parallel genomic assays». En: *Nature Reviews Genetics* 7.8 (ago. de 2006), págs. 632-644. DOI: 10.1038/nrg1901. URL: http://dx.doi.org/10.1038/nrg1901.

[115] Ranajit Chakraborty. *Hardy-Weinberg Equilibrium*. Jul. de 2005. DOI: 10.1002/0470011815.b2a05046. URL: http://dx.doi.org/10.1002/0470011815.b2a05046.

[116] Robert A. Power, Julian Parkhill y Tulio de Oliveira. «Microbial genome-wide association studies: lessons from human GWAS». En: *Nature Reviews Genetics* 18.1 (nov. de 2016), págs. 41-50. DOI: 10.1038/nrg.2016.132. URL: http://dx.doi.org/10.1038/nrg.2016.132.

[117] William S. Bush y Jason H. Moore. «Chapter 11: Genome-Wide Association Studies». En: *PLoS Computational Biology* 8.12 (dic. de 2012). Ed. por Fran Lewitter y Maricel Kann, e1002822. DOI: 10.1371/journal.pcbi.1002822. URL: http://dx.doi.org/10.1371/journal.pcbi.1002822.

[118] D. Altshuler, M. J. Daly y E. S. Lander. «Genetic Mapping in Human Disease». En: *Science* 322.5903 (nov. de 2008), págs. 881-888. DOI: 10.1126/science.1156409. URL: http://dx.doi.org/10.1126/science.1156409.

[119] Michael C. Wu y col. «Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test». En: *The American Journal of Human Genetics* 89.1 (jul. de 2011), págs. 82-93. DOI: 10.1016/j.ajhg.2011.05.029. URL: http://dx.doi.org/10.1016/j.ajhg.2011.05.029.

[120] Stephan Morgenthaler y William G. Thilly. «A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)». En: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615.1-2 (feb. de 2007), págs. 28-56. DOI: 10.1016/j.mrfmmm.2006.09.003. URL: http://dx.doi.org/10.1016/j.mrfmmm.2006.09.003.

[121] Andrew P. Morris y Eleftheria Zeggini. «An evaluation of statistical approaches to rare variant analysis in genetic association studies». En: *Genetic Epidemiology* 34.2 (oct. de 2009), págs. 188-193. DOI: 10.1002/gepi.20450. URL: http://dx.doi.org/10.1002/gepi.20450.

[122] Benjamin M. Neale y col. «Testing for an Unusual Distribution of Rare Variants». En: *PLoS Genetics* 7.3 (mar. de 2011). Ed. por Suzanne M. Leal, e1001322. DOI: 10.1371/journal.pgen.1001322. URL: http://dx.doi.org/10.1371/journal.pgen.1001322.

[123] Carl A Anderson y col. «Data quality control in genetic case-control association studies». En: *Nature Protocols* 5.9 (ago. de 2010), págs. 1564-1573. DOI: 10.1038/nprot.2010.116. URL: http://dx.doi.org/10.1038/nprot.2010.116.

[124] Jonathan Marchini y Bryan Howie. «Genotype imputation for genome-wide association studies». En: *Nature Reviews Genetics* 11.7 (jun. de 2010), págs. 499-511. DOI: 10.1038/nrg2796. URL: http://dx.doi.org/10.1038/nrg2796.

[125] B Devlin, Kathryn Roeder y Larry Wasserman. «Genomic Control, a New Approach to Genetic-Based Association Studies». En: *Theoretical Population Biology* 60.3 (nov. de 2001), págs. 155-166. DOI: 10.1006/tpbi.2001.1542. URL: http://dx.doi.org/10.1006/tpbi.2001.1542.

[126] Christoph Lippert y col. «FaST linear mixed models for genome-wide association studies». En: *Nature Methods* 8.10 (sep. de 2011), págs. 833-835. DOI: 10.1038/nmeth.1681. URL: http://dx.doi.org/10.1038/nmeth.1681.

[127] Christian Widmer y col. «Further Improvements to Linear Mixed Models for Genome-Wide Association Studies». En: *Scientific Reports* 4.1 (nov. de 2014). DOI: 10.1038/srep06874. URL: http://dx.doi.org/10.1038/srep06874.

[128] M. Kanehisa. «KEGG: Kyoto Encyclopedia of Genes and Genomes». En: *Nucleic Acids Research* 28.1 (ene. de 2000), págs. 27-30. DOI: 10.1093/nar/28.1.27. URL: http://dx.doi.org/10.1093/nar/28.1.27.

[129] Minoru Kanehisa. «Toward understanding the origin and evolution of cellular organisms». En: *Protein Science* 28.11 (sep. de 2019), págs. 1947-1951. DOI: 10.1002/pro.3715. URL: http://dx.doi.org/10.1002/pro.3715.

[130] Minoru Kanehisa y col. «KEGG: integrating viruses and cellular organisms». En: *Nucleic Acids Research* 49.D1 (oct. de 2020), págs. D545-D551. DOI: 10.1093/nar/gkaa970. URL: http://dx.doi.org/10.1093/nar/gkaa970.

[131] Dotun Ogunyemi y col. «Differentially expressed genes in adipocytokine signaling pathway of adipose tissue in pregnancy». En: *Journal of Diabetes Mellitus* 03.02 (2013), 86–95. DOI: 10.4236/jdm.2013.32013. URL: http://dx.doi.org/10.4236/jdm.2013.32013.

[132] Clare J. Wilhelm y col. «Adipocytokine signaling is altered in flinders sensitive line rats, and adiponectin correlates in humans with some symptoms of depression». En: *Pharmacology Biochemistry and Behavior* 103.3 (2013), págs. 643-651. DOI: 10.1016/j.pbb.2012.11.001. URL: http://dx.doi.org/10.1016/j.pbb.2012.11.001.

[133] Min Zhao, XiaoMo Li y Hong Qu. «EDdb: A web resource for eating disorder and its application to identify an extended adipocytokine signaling pathway related to eating disorder». En: *Science China Life Sciences* 56.12 (2013), págs. 1086-1096. DOI: 10.1007/s11427-013-4573-2. URL: http://dx.doi.org/10.1007/s11427-013-4573-2.

[134] David Carling. «AMPK signalling in health and disease». En: *Current Opinion in Cell Biology* 45 (abr. de 2017), págs. 31-37. DOI: 10.1016/j.ceb.2017.01.005. URL: http://dx.doi.org/10.1016/j.ceb.2017.01.005.

[135] Maria M. Mihaylova y Reuben J. Shaw. «The AMPK signalling pathway coordinates cell growth, autophagy and metabolism». En: *Nature Cell Biology* 13.9 (2011), págs. 1016-1023. DOI: 10.1038/ncb2329. URL: http://dx.doi.org/10.1038/ncb2329.

[136] David B. Shackelford y Reuben J. Shaw. «The LKB1–AMPK pathway: metabolism and growth control in tumour suppression». En: *Nature Reviews Cancer* 9.8 (ago. de 2009), págs. 563-575. DOI: 10.1038/nrc2676. URL: http://dx.doi.org/10.1038/nrc2676.

[137] Weitao Ji y col. «Mechanism of KLF4 Protection against Acute Liver Injury via Inhibition of Apelin Signaling». En: *Oxidative Medicine and Cellular Longevity* 2019 (oct. de 2019), págs. 1-10. DOI: 10.1155/2019/6140360. URL: http://dx.doi.org/10.1155/2019/6140360.

[138] Natalie O. Karpinich y Kathleen M. Caron. «Apelin Signaling». En: *Arteriosclerosis, Thrombosis, and Vascular Biology* 34.2 (feb. de 2014), págs. 239-241. DOI: 10.1161/atvbaha.113.302905. URL: http://dx.doi.org/10.1161/ATVBAHA.113.302905.

[139] Stacy M. Yadava y col. «miR-15b-5p promotes expression of proinflammatory cytokines in human placenta by inhibiting Apelin signaling pathway». En: *Placenta* 104 (ene. de 2021), págs. 8-15. DOI: 10.1016/j.placenta.2020.11.002. URL: http://dx.doi.org/10.1016/j.placenta.2020.11.002.

[140] Sofia La Vecchia y Carlos Sebastián. «Metabolic pathways regulating colorectal cancer initiation and progression». En: *Seminars in Cell Developmental Biology* 98 (feb. de 2020), págs. 63-70. DOI: 10.1016/j.semcdb.2019.05.018. URL: http://dx.doi.org/10.1016/j.semcdb.2019.05.018.

[141] Elena Puerta-García, Marisa Cañadas-Garre y Miguel Ángel Calleja-Hernández. «Molecular biomarkers in colorectal carcinoma». En: *Pharmacogenomics* 16.10 (jul. de 2015), págs. 1189-1222. DOI: 10.2217/pgs.15.63. URL: http://dx.doi.org/10.2217/PGS.15.63.

[142] *Glucagon signaling pathway - creative diagnostics*. https://www.creative-diagnostics.com/glucagon-signaling-pathway.htm. Accessed: 2021-6-16.

[143] *Glucagon Signaling Pathway- CUSABIO*. https://www.cusabio.com/pathway/Glucagon-signaling-pathway.html. Accessed: 2021-6-16.

[144] Sreedevi Chandrasekaran y Danail Bonchev. «Network analysis of human post-mortem microarrays reveals novel genes, microRNAs, and mechanistic scenarios of potential importance in fighting huntington's disease». En: *Computational and Structural Biotechnology Journal* 14 (2016), págs. 117-130. DOI: 10.1016/j.csbj.2016.02.001. URL: http://dx.doi.org/10.1016/j.csbj.2016.02.001.

[145] John Labbadia y Richard I. Morimoto. «Huntington's disease: underlying molecular mechanisms and emerging concepts». En: *Trends in Biochemical Sciences* 38.8 (ago. de 2013), págs. 378-385. DOI: 10.1016/j.tibs.2013.05.003. URL: http://dx.doi.org/10.1016/j.tibs.2013.05.003.

[146] Chiranjib Chakraborty y col. «Influence of miRNA in insulin signaling pathway and insulin resistance: micro-molecules with a major role in type-2 diabetes». En: *Wiley Interdisciplinary Reviews: RNA* 5.5 (jun. de 2014), págs. 697-712. DOI: 10.1002/wrna.1240. URL: http://dx.doi.org/10.1002/wrna.1240.

[147] Yvonne Ng, Georg Ramm y David E. James. «Dissecting the Mechanism of Insulin Resistance Using a Novel Heterodimerization Strategy to Activate Akt». En: *Journal of Biological Chemistry* 285.8 (feb. de 2010), págs. 5232-5239. DOI: 10.1074/jbc.m109.060632. URL: http://dx.doi.org/10.1074/jbc.M109.060632.

[148] Haroon y col. «Transcriptomic evidence that insulin signalling pathway regulates the ageing of subterranean termite castes». En: *Scientific Reports* 10.1 (mayo de 2020). DOI: 10.1038/s41598-020-64890-9. URL: http://dx.doi.org/10.1038/s41598-020-64890-9.

[149] Sreesha R. Sudhakar y col. «Insulin signalling elicits hunger-induced feeding in Drosophila». En: *Developmental Biology* 459.2 (mar. de 2020), págs. 87-99. DOI: 10.1016/j.ydbio.2019.11.013. URL: http://dx.doi.org/10.1016/j.ydbio.2019.11.013.

[150] Huaisha Xu y col. «Involvement of insulin signalling pathway in methamphetamine-induced hyperphosphorylation of Tau». En: *Toxicology* 408 (sep. de 2018), págs. 88-94. DOI: 10.1016/j.tox.2018.07.002. URL: http://dx.doi.org/10.1016/j.tox.2018.07.002.

[151] Yue Quan y col. «Mitochondrial ROS-Modulated mtDNA: A Potential Target for Cardiac Aging». En: *Oxidative Medicine and Cellular Longevity* 2020 (mar. de 2020), págs. 1-11. DOI: 10.1155/2020/9423593. URL: http://dx.doi.org/10.1155/2020/9423593.

[152] Francisca Salas-Pérez y col. «DNA methylation in genes of longevity-regulating pathways: association with obesity and metabolic complications». En: *Aging* 11.6 (mar. de 2019), 1874–1899. DOI: 10.18632/aging.101882. URL: http://dx.doi.org/10.18632/aging.101882.

[153] William R. Swindell y col. «Transcriptional profiling identifies strain-specific effects of caloric restriction and opposite responses in human and mouse white adipose tissue». En: *Aging* 10.4 (abr. de 2018), págs. 701-746. DOI: 10.18632/aging.101424. URL: http://dx.doi.org/10.18632/aging.101424.

[154] Celia Tengan, Gabriela Rodrigues y Rosely Godinho. «Nitric Oxide in Skeletal Muscle: Role on Mitochondrial Biogenesis and Function». En: *International Journal of Molecular Sciences* 13.12 (dic. de 2012), págs. 17160-17184. DOI: 10.3390/ijms131217160. URL: http://dx.doi.org/10.3390/ijms131217160.

[155] Virgil L. Anderson y Robert A. McLean. *Design of Experiments*. CRC Press, dic. de 2018. DOI: 10.1201/9781315141039. URL: http://dx.doi.org/10.1201/9781315141039.

[156] Živorad R. Lazić. *Design of Experiments in Chemical Engineering*. Sep. de 2004. DOI: 10.1002/3527604162. URL: http://dx.doi.org/10.1002/3527604162.

[157] Peter Goos y Bradley Jones. *Optimal Design of Experiments*. John Wiley Sons, Ltd, jul. de 2011. DOI: 10.1002/9781119974017. URL: http://dx.doi.org/10.1002/9781119974017.

[158] Messias Borges Silva. *Design of Experiments - Applications*. InTech, jun. de 2013. DOI: 10.5772/45728. URL: http://dx.doi.org/10.5772/45728.

[159] Fatima K. Suleiman, Kaihsiang Lin y Kyle J. Daun. «Development of a multivariate spectral emissivity model for an advanced high strength steel alloy through factorial design-of-experiments». En: *Journal of Quantitative Spectroscopy and Radiative Transfer* 271 (sep. de 2021), pág. 107693. DOI: 10.1016/j.jqsrt.2021.107693. URL: http://dx.doi.org/10.1016/j.jqsrt.2021.107693.

[160] Arman Chananipoor y col. «Optimization of the thermal performance of nano-encapsulated phase change material slurry in double pipe heat exchanger: Design of experiments using response surface methodology (RSM)». En: *Journal of Building Engineering* 34 (feb. de 2021), pág. 101929. DOI: 10.1016/j.jobe.2020.101929. URL: http://dx.doi.org/10.1016/j.jobe.2020.101929.

[161] Mebratu Tufa y col. «Study of sand-plastic composite using optimal mixture design of experiments for best compressive strength». En: *Materials Today: Proceedings* (mayo de 2021). DOI: 10.1016/j.matpr.2021.05.031. URL: http://dx.doi.org/10.1016/j.matpr.2021.05.031.

[162] Sushant S. Garud, Iftekhar A. Karimi y Markus Kraft. «Design of computer experiments: A review». En: *Computers Chemical Engineering* 106 (nov. de 2017), págs. 71-95. DOI: 10.1016/j.compchemeng.2017.05.010. URL: http://dx.doi.org/10.1016/j.compchemeng.2017.05.010.

[163] Milad Karimshoushtari y Carlo Novara. «Design of experiments for nonlinear system identification: A set membership approach». En: *Automatica* 119 (sep. de 2020), pág. 109036. DOI: 10.1016/j.automatica.2020.109036. URL: http://dx.doi.org/10.1016/j.automatica.2020.109036.

[164] Charlie Vanaret y col. «Two-phase approaches to optimal model-based design of experiments: how many experiments and which ones?» En: *Computers Chemical Engineering* 146 (mar. de 2021), pág. 107218. DOI: 10.1016/j.compchemeng.2020.107218. URL: http://dx.doi.org/10.1016/j.compchemeng.2020.107218.

[165] I.A. Choudhury y M.A. El-Baradie. «Surface roughness prediction in the turning of high-strength steel by factorial design of experiments». En: *Journal of Materials Processing Technology* 67.1-3 (mayo de 1997), págs. 55-61. DOI: 10.1016/s0924-0136(96)02818-x. URL: http://dx.doi.org/10.1016/S0924-0136(96)02818-X.

[166] Xiaochuan Tang y col. «Interaction-based feature selection using Factorial Design». En: *Neurocomputing* 281 (mar. de 2018), págs. 47-54. DOI: 10.1016/j.neucom.2017.11.058. URL: http://dx.doi.org/10.1016/j.neucom.2017.11.058.

[167] G Vicente y col. «Application of the factorial design of experiments and response surface methodology to optimize biodiesel production». En: *Industrial Crops and Products* 8.1 (mar. de 1998), págs. 29-35. DOI: 10.1016/s0926-6690(97)10003-6. URL: http://dx.doi.org/10.1016/S0926-6690(97)10003-6.

[168] N. R. Draper y F. Pukelsheim. «An overview of design of experiments». En: *Statistical Papers* 37.1 (mar. de 1996), págs. 1-32. DOI: 10.1007/bf02926157. URL: http://dx.doi.org/10.1007/BF02926157.

[169] Angelo Tulumello y J.D. Tulumello. «Yates' method analysis of 2n factorial design of experiments using the TI-59, for n = 3,4,5,6». En: *Computers Chemistry* 5.1 (ene. de 1981), págs. 55-66. DOI: 10.1016/0097-8485(81)80008-3. URL: http://dx.doi.org/10.1016/0097-8485(81)80008-3.

[170] Eva María Artime Ríos y col. «Genetic algorithm based on support vector machines for computer vision syndrome classification in health personnel». En: *Neural Computing and Applications* 32.5 (jun. de 2018), págs. 1239-1248. DOI: 10.1007/s00521-018-3581-3. URL: http://dx.doi.org/10.1007/s00521-018-3581-3.

[171] Eva María Artime Ríos y col. «Prediction of Computer Vision Syndrome in Health Personnel by Means of Genetic Algorithms and Binary Regression Trees». En: *Sensors* 19.12 (jun. de 2019), pág. 2800. DOI: 10.3390/s19122800. URL: http://dx.doi.org/10.3390/s19122800.

[172] P.J. García Nieto y col. «Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain)». En: *Environmental Research* 122 (abr. de 2013), págs. 1-10. DOI: 10.1016/j.envres.2013.01.001. URL: http://dx.doi.org/10.1016/j.envres.2013.01.001.

[173] Celestino Ordóñez Galán y col. «Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions». En: *Journal of Computational and Applied Mathematics* 311 (feb. de 2017), págs. 704-717. DOI: 10.1016/j.cam.2016.08.012. URL: http://dx.doi.org/10.1016/j.cam.2016.08.012.

[174] Juan Enrique Sánchez Lasheras y col. «Hybrid algorithm for the classification of prostate cancer patients of the MCC-Spain study based on support vector machines and genetic algorithms». En: *Neurocomputing* 452 (sep. de 2021), págs. 386-394. DOI: 10.1016/j.neucom.2019.08.113. URL: http://dx.doi.org/10.1016/j.neucom.2019.08.113.

[175] Ekain Azketa y col. «Algoritmo genético permutacional para el despliegue y la planificación de sistemas de tiempo real distribuidos». En: *Revista Iberoamericana de Automática e Informática Industrial RIAI* 10.3 (2013), págs. 344-355. DOI: 10.1016/j.riai.2013.05.006. URL: http://dx.doi.org/10.1016/j.riai.2013.05.006.

[176] Julieta Sol Dussaut y col. «Algoritmos Evolutivos Multiobjetivo aplicados a la Selección de Características en Microarrays de Datos de Cáncer». En: *Entre ciencia e ingeniería* 14.28 (dic. de 2020), págs. 40-45. DOI: 10.31908/19098367.2014. URL: http://dx.doi.org/10.31908/19098367.2014.

[177] C. Fernández y col. «Optimización de Parámetros Utilizando los Métodos de Monte Carlo y Algoritmos Evolutivos. Aplicación a un Controlador de Seguimiento de Trayectoria en Sistemas no Lineales». En: *Revista Iberoamericana de Automática e Informática industrial* 16.1 (dic. de 2018), pág. 89. DOI: 10.4995/riai.2018.8796. URL: http://dx.doi.org/10.4995/riai.2018.8796.

[178] Carlos García, Edwin García y Fernando Villada. «Algoritmo Evolutivo Eficiente Aplicado a la Planeación de la Expansión de Sistemas de Distribución». En: *Información tecnológica* 23.4 (2012), págs. 3-10. DOI: 10.4067/s0718-07642012000400002. URL: http://dx.doi.org/10.4067/S0718-07642012000400002.

[179] Miguel Jiménez-Carrión. «Algoritmo Genético Simple para Resolver el Problema de Programación de la Tienda de Trabajo (Job Shop Scheduling)». En: *Información tecnológica* 29.5 (oct. de 2018), págs. 299-314. DOI: 10.4067/s0718-07642018000500299. URL: http://dx.doi.org/10.4067/S0718-07642018000500299.

[180] María de los Ángeles Rodríguez-Cevallos y col. «Implementación de un algoritmo genético mediante una aplicación informática basado en la computación neuronal y evolutiva para obtener el cromosoma mejor adaptado». En: *Revista Boletín Redipe* 9.8 (ago. de 2020), págs. 116-131. DOI: 10.36260/rbr.v9i8.1045. URL: http://dx.doi.org/10.36260/rbr.v9i8.1045.

[181] D. E. Goldberg y K. Deb. «A comparative analysis of selection schemes used in genetic algorithms». En: *Foundations of Genetic Algorithms* 1 (1991), págs. 69-93.

[182] Z. Wang y A. Sobey. «A comparative review between Genetic Algorithm use in composite optimisation and the state-of-the-art in evolutionary computation». En: *Composite Structures* 233.11173 (2020), pág. 9.

[183] M. De la Maza y B. Tidor. «An analysis of selection procedures with particular attention paid to proportional and Boltzmann selection». En: *En Forrest*. Proceedings of the Fifth International Conference on Genetic Algorithms. San Mateo: Morgan Kaufman, 1993, págs. 124-131.

[184] Shengxiang Yang. «Genetic Algorithms with Memory- and Elitism-Based Immigrants in Dynamic Environments». En: *Evolutionary Computation* 16.3 (sep. de 2008), págs. 385-416. DOI: 10.1162/evco.2008.16.3.385. URL: http://dx.doi.org/10.1162/evco.2008.16.3.385.

[185] Ahmad Hassanat y col. «Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach». En: *Information* 10.12 (dic. de 2019), pág. 390. DOI: 10.3390/info10120390. URL: http://dx.doi.org/10.3390/info10120390.

[186] Mustafa Kaya. «The effects of two new crossover operators on genetic algorithm performance». En: *Applied Soft Computing* 11.1 (ene. de 2011), págs. 881-890. DOI: 10.1016/j.asoc.2010.01.008. URL: http://dx.doi.org/10.1016/j.asoc.2010.01.008.

[187] X. B. Hu y E. Di Paolo. «An efficient Genetic Algorithm with uniform crossover for the multi-objective Airport Gate Assignment Problem». En: *2007 IEEE Congress on Evolutionary Computation*. IEEE, sep. de 2007. DOI: 10.1109/cec.2007.4424454. URL: http://dx.doi.org/10.1109/CEC.2007.4424454.

[188] W. M. Spears y K. A. De Jong. «On the virtues of parameterized uniform crossover». En: *En Belew*. Ed. por R. K. y L. B Booker. Proceedings of the Fourth International Conference on Genetic algorithms . San Mateo, California: Morgan Kaufmann, 1991, págs. 230-236.

[189] Xueqin Lü y col. «Energy management of hybrid electric vehicles: A review of energy optimization of fuel cell hybrid power system based on genetic algorithm». En: *Energy Conversion and Management* 205 (feb. de 2020), pág. 112474. DOI: 10.1016/j.enconman.2020.112474. URL: http://dx.doi.org/10.1016/j.enconman.2020.112474.

[190] ZhenZhou Wang y Adam Sobey. «A comparative review between Genetic Algorithm use in composite optimisation and the state-of-the-art in evolutionary computation». En: *Composite Structures* 233 (feb. de 2020), pág. 111739. DOI: 10.1016/j.compstruct.2019.111739. URL: http://dx.doi.org/10.1016/j.compstruct.2019.111739.

[191] Samille Santos Rocha y col. «Applying optimization algorithms for spatial estimation of travel demand variables». En: *Transportation Research Interdisciplinary Perspectives* 10 (2021), pág. 100369. DOI: 10.1016/j.trip.2021.100369. URL: http://dx.doi.org/10.1016/j.trip.2021.100369.

[192] Abhilash Singh, Sandeep Sharma y Jitendra Singh. «Nature-inspired algorithms for Wireless Sensor Networks: A comprehensive survey». En: *Computer Science Review* 39 (feb. de 2021), pág. 100342. DOI: 10.1016/j.cosrev.2020.100342. URL: http://dx.doi.org/10.1016/j.cosrev.2020.100342.

[193] Shaymaa Adnan Abdulrahman y col. «Comparative study for 8 computational intelligence algorithms for human identification». En: *Computer Science Review* 36 (mayo de 2020), pág. 100237. DOI: 10.1016/j.cosrev.2020.100237. URL: http://dx.doi.org/10.1016/j.cosrev.2020.100237.

[194] Biswanath Chowdhury y Gautam Garai. «A review on multiple sequence alignment from the perspective of genetic algorithm». En: *Genomics* 109.5-6 (oct. de 2017), págs. 419-431. DOI: 10.1016/j.ygeno.2017.06.007. URL: http://dx.doi.org/10.1016/j.ygeno.2017.06.007.

[195] Panpan Cai y col. «Parallel genetic algorithm based automatic path planning for crane lifting in complex environments». En: *Automation in Construction* 62 (feb. de 2016), págs. 133-147. DOI: 10.1016/j.autcon.2015.09.007. URL: http://dx.doi.org/10.1016/j.autcon.2015.09.007.

[196] Heinrich Niederhausen. «Factorials and Stirling numbers in the algebra of formal Laurent series II: zazb=t». En: *Discrete Mathematics* 132.1-3 (sep. de 1994), págs. 197-213. DOI: 10.1016/0012-365x(94)90238-0. URL: http://dx.doi.org/10.1016/0012-365X(94)90238-0.

[197] Arpan Kumar Kar. «Bio inspired computing – A review of algorithms and scope of applications». En: *Expert Systems with Applications* 59 (oct. de 2016), págs. 20-32. DOI: 10.1016/j.eswa.2016.04.018. URL: http://dx.doi.org/10.1016/j.eswa.2016.04.018.

[198] Mahdi Sedghi, Ali Ahmadian y Masoud Aliakbar-Golkar. «Assessment of optimization algorithms capability in distribution network planning: Review, comparison and modification techniques». En: *Renewable and Sustainable Energy Reviews* 66 (dic. de 2016), págs. 415-434. DOI: 10.1016/j.rser.2016.08.027. URL: http://dx.doi.org/10.1016/j.rser.2016.08.027.

[199] John H Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence*. Londres, Inglaterra: MIT Press, 1992.

[200] Lee Altenberg. «The Schema Theorem and Price's Theorem». En: *Foundations of Genetic Algorithms*. Elsevier, 1995, págs. 23-49. DOI: 10.1016/b978-1-55860-356-1.50006-6. URL: http://dx.doi.org/10.1016/B978-1-55860-356-1.50006-6.

[201] Riccardo Poli. «Recursive Conditional Schema Theorem, Convergence and Population Sizing in Genetic Algorithms». En: *Foundations of Genetic Algorithms 6*. Elsevier, 2001, págs. 143-163. DOI: 10.1016/b978-155860734-7/50091-3. URL: http://dx.doi.org/10.1016/B978-155860734-7/50091-3.

[202] H. Van Hove y A. Verschoren. «A fuzzy Schema Theorem». En: *Fuzzy Sets and Systems* 94.1 (feb. de 1998), págs. 93-99. DOI: 10.1016/s0165-0114(96)00210-2. URL: http://dx.doi.org/10.1016/S0165-0114(96)00210-2.

[203] Xiao Feng Yin y Li Pheng Khoo. «An exact schema theorem for adaptive genetic algorithm and its application to machine cell formation». En: *Expert Systems with Applications* 38.7 (jul. de 2011), págs. 8538-8552. DOI: 10.1016/j.eswa.2011.01.055. URL: http://dx.doi.org/10.1016/j.eswa.2011.01.055.

[204] Heinz Mühlenbein. «Evolution in Time and Space – The Parallel Genetic Algorithm». En: *Foundations of Genetic Algorithms*. Elsevier, 1991, págs. 316-337. DOI: 10.1016/b978-0-08-050684-5.50023-9. URL: http://dx.doi.org/10.1016/B978-0-08-050684-5.50023-9.

[205] Colin R. Reeves y Jonathan E. Rowe. *Genetic Algorithms—Principles and Perspectives*. Springer US, 2002. DOI: 10.1007/b101880. URL: http://dx.doi.org/10.1007/b101880.

[206] N. Radcliffe. «Forma Analysis and Random Respectful Recombination». En: *ICGA*. 1991.

[207] Michael D. Vose. *The simple genetic algorithm - foundations and theory*. Complex adaptive systems. MIT Press, 1999. ISBN: 978-0-262-22058-3.

[208] Ahmed Abdulaal y col. «Comparison of deep learning with regression analysis in creating predictive models for SARS-CoV-2 outcomes». En: 20 (nov. de 2020). DOI: 10.1186/s12911-020-01316-6. URL: http://dx.doi.org/10.1186/s12911-020-01316-6.

[209] Ellysia Jumin y col. «Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction». En: *Engineering Applications of Computational Fluid Mechanics* 14.1 (ene. de 2020), págs. 713-725. DOI: 10.1080/19942060.2020.1758792. URL: http://dx.doi.org/10.1080/19942060.2020.1758792.

[210] Shrikant I. Bangdiwala. «Regression: simple linear». En: *International Journal of Injury Control and Safety Promotion* 25.1 (ene. de 2018), págs. 113-115. DOI: 10.1080/17457300.2018.1426702. URL: http://dx.doi.org/10.1080/17457300.2018.1426702.

[211] B. Sch.ºlkopf, A. J. Smola y F. Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, Massachusetts: MIT Press, 2018.

[212] V. Vapnik y A. Chervonenkis. «A Note on One Class of Perceptrons». En: *Automation and Remote Control* 25 (1964), págs. 103-109.

[213] Jeonghyun Baek y Euntai Kim. «A new support vector machine with an optimal additive kernel». En: *Neurocomputing* 329 (feb. de 2019), págs. 279-299. DOI: 10.1016/j.neucom.2018.10.032. URL: http://dx.doi.org/10.1016/j.neucom.2018.10.032.

[214] Anirban Chatterjee, Kelly Fermoyle y Padma Raghavan. «Characterizing sparse preconditioner performance for the support vector machine kernel». En: *Procedia Computer Science* 1.1 (mayo de 2010), págs. 367-375. DOI: 10.1016/j.procs.2010.04.040. URL: http://dx.doi.org/10.1016/j.procs.2010.04.040.

[215] Harris Drucker, Behzad Shahrary y David C Gibbon. «Support vector machines: relevance feedback and information retrieval». En: *Information Processing Management* 38.3 (mayo de 2002), págs. 305-323. DOI: 10.1016/s0306-4573(01)00037-1. URL: http://dx.doi.org/10.1016/S0306-4573(01)00037-1.

[216] D. Basu. «On Sampling with and Without Replacement». En: *Selected Works of Debabrata Basu*. Springer New York, dic. de 2010, págs. 73-80. DOI: 10.1007/978-1-4419-5825-9_17. URL: http://dx.doi.org/10.1007/978-1-4419-5825-9_17.

[217] Des Raj y Salem H. Khamis. «Some Remarks on Sampling with Replacement». En: *The Annals of Mathematical Statistics* 29.2 (jun. de 1958), págs. 550-557. DOI: 10.1214/aoms/1177706630. URL: http://dx.doi.org/10.1214/aoms/1177706630.

[218] Robert Trevethan. «Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice». En: *Frontiers in Public Health* 5 (nov. de 2017). DOI: 10.3389/fpubh.2017.00307. URL: http://dx.doi.org/10.3389/fpubh.2017.00307.

[219] Rajeev Kumar y Abhaya Indrayan. «Receiver operating characteristic (ROC) curve for medical researchers». En: *Indian Pediatrics* 48.4 (abr. de 2011), págs. 277-287. DOI: 10.1007/s13312-011-0055-4. URL: http://dx.doi.org/10.1007/s13312-011-0055-4.

[220] Kelly H. Zou, A. James O'Malley y Laura Mauri. «Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models». En: *Circulation* 115.5 (feb. de 2007), págs. 654-657. DOI: 10.1161/circulationaha.105.594929. URL: http://dx.doi.org/10.1161/CIRCULATIONAHA.105.594929.

[221] *Genome-Wide Association Studies and Genomic Prediction*. Humana Press, 2013. DOI: 10.1007/978-1-62703-447-0. URL: http://dx.doi.org/10.1007/978-1-62703-447-0.

[222] Marti J. Anderson y Pierre Legendre. «An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model». En: *Journal of Statistical Computation and Simulation* 62.3 (feb. de 1999), págs. 271-303. DOI: 10.1080/00949659908811936. URL: http://dx.doi.org/10.1080/00949659908811936.

[223] Marco Marozzi. «A bi-aspect nonparametric test for the two-sample location problem». En: *Computational Statistics Data Analysis* 44.4 (ene. de 2004), págs. 639-648. DOI: 10.1016/s0167-9473(02)00279-7. URL: http://dx.doi.org/10.1016/S0167-9473(02)00279-7.

[224] Bill Shipley. En: *Statistics and Computing* 10.3 (2000), págs. 253-257. DOI: 10.1023/a:1008943611855. URL: http://dx.doi.org/10.1023/A:1008943611855.

[225] Michael D. Ernst y William R. Schucany. «A Class of Permutation Tests of Bivariate Interchangeability». En: *Journal of the American Statistical Association* 94.445 (mar. de 1999), págs. 273-284. DOI: 10.1080/01621459.1999.10473843. URL: http://dx.doi.org/10.1080/01621459.1999.10473843.

[226] Susan Dadakis Horn. «Goodness-of-Fit Tests for Discrete Data: A Review and an Application to a Health Impairment Scale». En: *Biometrics* 33.1 (mar. de 1977), pág. 237. DOI: 10.2307/2529319. URL: http://dx.doi.org/10.2307/2529319.

[227] Andrew R. Craig y Wayne W. Fisher. «Randomization tests as alternative analysis methods for behavior-analytic data». En: *Journal of the Experimental Analysis of Behavior* 111.2 (feb. de 2019), págs. 309-328. DOI: 10.1002/jeab.500. URL: http://dx.doi.org/10.1002/jeab.500.

[228] M. Marozzi. «Some remarks about the number of permutations one should consider to perform a permutation test». En: *Statistica; Vol 64* (2007), No 1 (2004); 193-201. DOI: 10.6092/ISSN.1973-2201/32. URL: http://rivista-statistica.unibo.it/article/view/32.

[229] Brian L Browning. «PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies». En: *BMC Bioinformatics* 9.1 (jul. de 2008). DOI: 10.1186/1471-2105-9-309. URL: http://dx.doi.org/10.1186/1471-2105-9-309.

[230] Min Dong y Yan Wu. «Dynamic Crossover and Mutation Genetic Algorithm Based on Expansion Sampling». En: *Artificial Intelligence and Computational Intelligence*. Springer Berlin Heidelberg, 2009, págs. 141-149. DOI: 10.1007/978-3-642-05253-8_16. URL: http://dx.doi.org/10.1007/978-3-642-05253-8_16.

[231] Sourabh Katoch, Sumit Singh Chauhan y Vijay Kumar. «A review on genetic algorithm: past, present, and future». En: *Multimedia Tools and Applications* 80.5 (oct. de 2020), págs. 8091-8126. DOI: 10.1007/s11042-020-10139-6. URL: http://dx.doi.org/10.1007/s11042-020-10139-6.

[232] Carlos M. Fernandes y col. «A Study on the Mutation Rates of a Genetic Algorithm Interacting with a Sandpile». En: *Applications of Evolutionary Computation*. Springer Berlin Heidelberg, 2011, págs. 32-42. DOI: 10.1007/978-3-642-20525-5_4. URL: http://dx.doi.org/10.1007/978-3-642-20525-5_4.

[233] Yu-an Zhang, Makoto Sakamoto e Hiroshi Furutani. «Effects of Population Size and Mutation Rate on Results of Genetic Algorithm». En: *2008 Fourth International Conference on Natural Computation*. IEEE, 2008. DOI: 10.1109/icnc.2008.345. URL: http://dx.doi.org/10.1109/ICNC.2008.345.

[234] Michael Frigge, David C. Hoaglin y Boris Iglewicz. «Some Implementations of the Boxplot». En: *The American Statistician* 43.1 (feb. de 1989), pág. 50. DOI: 10.2307/2685173. URL: http://dx.doi.org/10.2307/2685173.

[235] Fernando Marmolejo-Ramos y Tian Siva Tian. «The shifting boxplot. A boxplot based on essential summary statistics around the mean.» En: *International Journal of Psychological Research* 3.1 (jun. de 2010), 37–45. DOI: 10.21500/20112084.823. URL: http://dx.doi.org/10.21500/20112084.823.

[236] Joseph L Gage, Natalia de Leon y Murray K Clayton. «Comparing Genome-Wide Association Study Results from Different Measurements of an Underlying Phenotype». En: *G3 Genes|Genomes|Genetics* 8.11 (nov. de 2018), 3715–3722. DOI: 10.1534/g3.118.200700. URL: http://dx.doi.org/10.1534/g3.118.200700.

[237] Wei Jiang, Jing-Hao Xue y Weichuan Yu. «What is the probability of replicating a statistically significant association in genome-wide association studies?» En: *Briefings in Bioinformatics* (sep. de 2016), bbw091. DOI: 10.1093/bib/bbw091. URL: http://dx.doi.org/10.1093/bib/bbw091.

[238] Nathan D. Miller y col. «A robust, high-throughput method for computing maize ear, cob, and kernel attributes automatically from images». En: *The Plant Journal* 89.1 (nov. de 2016), 169–178. DOI: 10.1111/tpj.13320. URL: http://dx.doi.org/10.1111/tpj.13320.

[239] Jonas Patron y col. «Assessing the performance of genome-wide association studies for predicting disease risk». En: *PLOS ONE* 14.12 (dic. de 2019). Ed. por Joseph Devaney, e0220215. DOI: 10.1371/journal.pone.0220215. URL: http://dx.doi.org/10.1371/journal.pone.0220215.

[240] Minta Thomas y col. «Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk». En: *The American Journal of Human Genetics* 107.3 (sep. de 2020), págs. 432-444. DOI: 10.1016/j.ajhg.2020.07.006. URL: http://dx.doi.org/10.1016/j.ajhg.2020.07.006.

[241] Qishan Wang y col. «A SUPER Powerful Method for Genome Wide Association Study». En: *PLoS ONE* 9.9 (2014). Ed. por Yun Li, e107684. DOI: 10.1371/journal.pone.0107684. URL: http://dx.doi.org/10.1371/journal.pone.0107684.

[242] Mohamed M. Ali, Shane A. Phillips y Abeer M. Mahmoud. «HIF1/TET1 Pathway Mediates Hypoxia-Induced Adipocytokine Promoter Hypomethylation in Human Adipocytes». En: *Cells* 9.1 (ene. de 2020), pág. 134. DOI: 10.3390/cells9010134. URL: http://dx.doi.org/10.3390/cells9010134.

[243] Etan Orgel y Steven D. Mittelman. «The Links Between Insulin Resistance, Diabetes, and Cancer». En: *Current Diabetes Reports* 13.2 (2012), 213–222. DOI: 10.1007/s11892-012-0356-6. URL: http://dx.doi.org/10.1007/s11892-012-0356-6.

[244] Silvia Riondino. «Obesity and colorectal cancer: Role of adipokines in tumor initiation and progression». En: *World Journal of Gastroenterology* 20.18 (2014), pág. 5177. DOI: 10.3748/wjg.v20.i18.5177. URL: http://dx.doi.org/10.3748/wjg.v20.i18.5177.

[245] Yanjie Yang y col. «Apelin/APJ system and cancer». En: *Clinica Chimica Acta* 457 (jun. de 2016), págs. 112-116. DOI: 10.1016/j.cca.2016.04.001. URL: http://dx.doi.org/10.1016/j.cca.2016.04.001.

[246] Bing Guo y col. «AMPK promotes the survival of colorectal cancer stem cells». En: *Animal Models and Experimental Medicine* 1.2 (jun. de 2018), págs. 134-142. DOI: 10.1002/ame2.12016. URL: http://dx.doi.org/10.1002/ame2.12016.

[247] Amreen Mughal y Stephen T. O'Rourke. «Vascular effects of apelin: Mechanisms and therapeutic potential». En: *Pharmacology Therapeutics* 190 (oct. de 2018), págs. 139-147. DOI: 10.1016/j.pharmthera.2018.05.013. URL: http://dx.doi.org/10.1016/j.pharmthera.2018.05.013.

[248] François-Xavier Picault y col. «Tumour co-expression of apelin and its receptor is the basis of an autocrine loop involved in the growth of colon adenocarcinomas». En: *European Journal of Cancer* 50.3 (feb. de 2014), págs. 663-674. DOI: 10.1016/j.ejca.2013.11.017. URL: http://dx.doi.org/10.1016/j.ejca.2013.11.017.

[249] Podgórska y col. «Evaluation of Apelin and Apelin Receptor Level in the Primary Tumor and Serum of Colorectal Cancer Patients». En: *Journal of Clinical Medicine* 8.10 (sep. de 2019), pág. 1513. DOI: 10.3390/jcm8101513. URL: http://dx.doi.org/10.3390/jcm8101513.

[250] Karam S. Boparai y col. «A Serrated Colorectal Cancer Pathway Predominates over the Classic WNT Pathway in Patients with Hyperplastic Polyposis Syndrome». En: *The American Journal of Pathology* 178.6 (ene. de 2011), 2700–2707. DOI: 10.1016/j.ajpath.2011.02.023. URL: http://dx.doi.org/10.1016/j.ajpath.2011.02.023.

[251] J Christopher y col. «Straight-to-test for the two-week-wait colorectal cancer pathway under the updated NICE guidelines reduces time to cancer diagnosis and treatment». En: *The Annals of The Royal College of Surgeons of England* 101.5 (mayo de 2019), 333–339. DOI: 10.1308/rcsann.2019.0022. URL: http://dx.doi.org/10.1308/rcsann.2019.0022.

[252] W Maclean y col. «The two-week rule colorectal cancer pathway: an update on recent practice, the unsustainable burden on diagnostics and the role of faecal immunochemical testing». En: *The Annals of The Royal College of Surgeons of England* 102.4 (abr. de 2020), 308–311. DOI: 10.1308/rcsann.2020.0019. URL: http://dx.doi.org/10.1308/rcsann.2020.0019.

[253] Fengfeng Wang y col. «Multiple Regression Analysis of mRNA-miRNA Associations in Colorectal Cancer Pathway». En: *BioMed Research International* 2014 (2014), 1–8. DOI: 10.1155/2014/676724. URL: http://dx.doi.org/10.1155/2014/676724.

[254] Tzu-Wei Yang y col. «Enterotype-based Analysis of Gut Microbiota along the Conventional Adenoma-Carcinoma Colorectal Cancer Pathway». En: *Scientific Reports* 9.1 (jul. de 2019). DOI: 10.1038/s41598-019-45588-z. URL: http://dx.doi.org/10.1038/s41598-019-45588-z.

[255] Dalila Azzout-Marniche y col. «Liver glyconeogenesis: a pathway to cope with postprandial amino acid excess in high-protein fed rats?» En: *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 292.4 (abr. de 2007), R1400–R1407. DOI: 10.1152/ajpregu.00566.2006. URL: http://dx.doi.org/10.1152/ajpregu.00566.2006.

[256] Liangyou Rui. *Energy Metabolism in the Liver*. Ene. de 2014. DOI: 10.1002/cphy.c130024. URL: http://dx.doi.org/10.1002/cphy.c130024.

[257] Tiange Wang, Guang Ning y Zachary Bloomgarden. «Diabetes and cancer relationships». En: *Journal of Diabetes* 5.4 (jun. de 2013), 378–390. DOI: 10.1111/1753-0407.12057. URL: http://dx.doi.org/10.1111/1753-0407.12057.

[258] Takashi Yagi y col. «Glucagon promotes colon cancer cell growth via regulating AMPK and MAPK pathways». En: *Oncotarget* 9.12 (ene. de 2018), págs. 10650-10664. DOI: 10.18632/oncotarget.24367. URL: http://dx.doi.org/10.18632/oncotarget.24367.

[259] Zhuoxuan Wu y col. «Analysis of potential genes and pathways associated with the colorectal normal mucosa-adenoma-carcinoma sequence». En: *Cancer Medicine* 7.6 (abr. de 2018), págs. 2555-2566. DOI: 10.1002/cam4.1484. URL: http://dx.doi.org/10.1002/cam4.1484.

[260] Devin Abrahami y col. «Incretin-based Drugs and the Incidence of Colorectal Cancer in Patients with Type 2 Diabetes». En: *Epidemiology* 29.2 (mar. de 2018), 246–253. DOI: 10.1097/ede.0000000000000793. URL: http://dx.doi.org/10.1097/EDE.0000000000000793.

[261] E. Danielle Dean y col. «Interrupted Glucagon Signaling Reveals Hepatic Cell Axis and Role for L-Glutamine in Cell Proliferation». En: *Cell Metabolism* 25.6 (jun. de 2017), 1362-1373.e5. DOI: 10.1016/j.cmet.2017.05.011. URL: http://dx.doi.org/10.1016/j.cmet.2017.05.011.

[262] Carol J. Lam y col. «Glucagon Receptor Antagonist–Stimulated -Cell Proliferation Is Severely Restricted With Advanced Age». En: *Diabetes* 68.5 (mar. de 2019), 963–974. DOI: 10.2337/db18-1293. URL: http://dx.doi.org/10.2337/db18-1293.

[263] Ranka Kanda y col. «Expression of the glucagon-like peptide-1 receptor and its role in regulating autophagy in endometrial cancer». En: *BMC Cancer* 18.1 (jun. de 2018). DOI: 10.1186/s12885-018-4570-8. URL: http://dx.doi.org/10.1186/s12885-018-4570-8.

[264] S. N. Illarioshkin y col. «Molecular Pathogenesis in Huntington's Disease». En: *Biochemistry (Moscow)* 83.9 (sep. de 2018), 1030–1039. DOI: 10.1134/s0006297918090043. URL: http://dx.doi.org/10.1134/S0006297918090043.

[265] Praveen Dayalu y Roger L. Albin. «Huntington Disease». En: *Neurologic Clinics* 33.1 (feb. de 2015), 101–114. DOI: 10.1016/j.ncl.2014.09.003. URL: http://dx.doi.org/10.1016/j.ncl.2014.09.003.

[266] Samuel Frank. «Treatment of Huntington's Disease». En: *Neurotherapeutics* 11.1 (dic. de 2013), 153–160. DOI: 10.1007/s13311-013-0244-z. URL: http://dx.doi.org/10.1007/s13311-013-0244-z.

[267] Jussi O.T. Sipilä y col. «Epidemiology of Huntington's disease in Finland». En: *Parkinsonism Related Disorders* 21.1 (ene. de 2015), 46–49. DOI: 10.1016/j.parkreldis.2014.10.025. URL: http://dx.doi.org/10.1016/j.parkreldis.2014.10.025.

[268] Robert J. Ferrante y col. «Cytochrome C and Caspase-9 Expression in Huntington 's Disease». En: *NeuroMolecular Medicine* 1.3 (2002), págs. 183-196. DOI: 10.1385/nmm:1:3:183. URL: http://dx.doi.org/10.1385/NMM:1:3:183.

[269] Yi-Min Sun, Yan-Bin Zhang y Zhi-Ying Wu. «Huntington's Disease: Relationship Between Phenotype and Genotype». En: *Molecular Neurobiology* 54.1 (ene. de 2016), págs. 342-348. DOI: 10.1007/s12035-015-9662-8. URL: http://dx.doi.org/10.1007/s12035-015-9662-8.

[270] Yu-Fen Huang, Hsiang-Yuan Yeh y Von-Wun Soo. «Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation». En: *BMC Medical Genomics* 6.Suppl 3 (2013), S4. DOI: 10.1186/1755-8794-6-s3-s4. URL: http://dx.doi.org/10.1186/1755-8794-6-S3-S4.

[271] Silke Crommen y Marie-Christine Simon. «Microbial Regulation of Glucose Metabolism and Insulin Resistance». En: *Genes* 9.1 (dic. de 2017), pág. 10. DOI: 10.3390/genes9010010. URL: http://dx.doi.org/10.3390/genes9010010.

[272] Gloria González-Saldivar y col. «Skin Manifestations of Insulin Resistance: From a Biochemical Stance to a Clinical Diagnosis and Management». En: *Dermatology and Therapy* 7.1 (dic. de 2016), 37–51. DOI: 10.1007/s13555-016-0160-3. URL: http://dx.doi.org/10.1007/s13555-016-0160-3.

[273] Ahmad Al-Mrabeh y col. «Morphology of the pancreas in type 2 diabetes: effect of weight loss with or without normalisation of insulin secretory capacity». En: *Diabetologia* 59.8 (mayo de 2016), 1753–1759. DOI: 10.1007/s00125-016-3984-6. URL: http://dx.doi.org/10.1007/s00125-016-3984-6.

[274] Juan Antonio Paniagua. «Nutrition, insulin resistance and dysfunctional adipose tissue determine the different components of metabolic syndrome». En: *World Journal of Diabetes* 7.19 (2016), pág. 483. DOI: 10.4239/wjd.v7.i19.483. URL: http://dx.doi.org/10.4239/wjd.v7.i19.483.

[275] Yohannes Tsegyie Wondmkun. « Obesity, Insulin Resistance, and Type 2 Diabetes: Associations and Therapeutic Implications ». En: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* Volume 13 (oct. de 2020), 3611–3616. DOI: 10.2147/dmso.s275898. URL: http://dx.doi.org/10.2147/DMSO.S275898.

[276] Audrey E. Brown y Mark Walker. «Genetics of Insulin Resistance and the Metabolic Syndrome». En: *Current Cardiology Reports* 18.8 (jun. de 2016). DOI: 10.1007/s11886-016-0755-4. URL: http://dx.doi.org/10.1007/s11886-016-0755-4.

[277] Sona Kang, Linus T-Y. Tsai y Evan D. Rosen. «Nuclear Mechanisms of Insulin Resistance». En: *Trends in Cell Biology* 26.5 (mayo de 2016), 341–351. DOI: 10.1016/j.tcb.2016.01.002. URL: http://dx.doi.org/10.1016/j.tcb.2016.01.002.

[278] N. Fernando Carrasco, F. José Eduardo Galgani y J. Marcela Reyes. «Síndrome de resistencia a la insulina. estudio y manejo». En: *Revista Médica Clínica Las Condes* 24.5 (sep. de 2013), 827–837. DOI: 10.1016/s0716-8640(13)70230-x. URL: http://dx.doi.org/10.1016/S0716-8640(13)70230-X.

[279] Felipe Pollak y col. «II Consenso de la Sociedad Chilena de Endocrinología y Diabetes sobre resistencia a la insulina». En: *Revista médica de Chile* 143.5 (mayo de 2015), 627–636. DOI: 10.4067/s0034-98872015000500012. URL: http://dx.doi.org/10.4067/S0034-98872015000500012.

[280] Max C. Petersen y Gerald I. Shulman. «Mechanisms of Insulin Action and Insulin Resistance». En: *Physiological Reviews* 98.4 (2018), 2133–2223. DOI: 10.1152/physrev.00063.2017. URL: http://dx.doi.org/10.1152/physrev.00063.2017.

[281] Alexandre A. da Silva y col. «Role of Hyperinsulinemia and Insulin Resistance in Hypertension: Metabolic Syndrome Revisited». En: *Canadian Journal of Cardiology* 36.5 (mayo de 2020), 671–682. DOI: 10.1016/j.cjca.2020.02.066. URL: http://dx.doi.org/10.1016/j.cjca.2020.02.066.

[282] Sonali Pechlivanis y col. «Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect». En: *Endocrine-Related Cancer* 14.3 (sep. de 2007), págs. 733-740. DOI: 10.1677/erc-07-0107. URL: http://dx.doi.org/10.1677/ERC-07-0107.

[283] Su Yon Jung y col. «Genetic variants and traits related to insulin-like growth factor-I and insulin resistance and their interaction with lifestyles on postmenopausal colorectal cancer risk». En: *PLOS ONE* 12.10 (oct. de 2017). Ed. por Aamir Ahmad, e0186296. DOI: 10.1371/journal.pone.0186296. URL: http://dx.doi.org/10.1371/journal.pone.0186296.

[284] Ana Elisa Lohmann y col. «Association of Obesity-Related Metabolic Disruptions With Cancer Risk and Outcome». En: *Journal of Clinical Oncology* 34.35 (dic. de 2016), págs. 4249-4255. DOI: 10.1200/jco.2016.69.6187. URL: http://dx.doi.org/10.1200/JCO.2016.69.6187.

[285] Y Poloz y V Stambolic. «Obesity and cancer, a case for insulin signaling». En: *Cell Death Disease* 6.12 (dic. de 2015), e2037-e2037. DOI: 10.1038/cddis.2015.381. URL: http://dx.doi.org/10.1038/cddis.2015.381.

[286] W. Robert Bruce, Thomas M. S. Wolever y Adria Giacca. «Mechanisms Linking Diet and Colorectal Cancer: The Possible Role of Insulin Resistance». En: *Nutrition and Cancer* 37.1 (mayo de 2000), 19–26. DOI: 10.1207/s15327914nc3701_2. URL: http://dx.doi.org/10.1207/S15327914NC3701_2.

[287] Francesca Cirillo y col. «Obesity, Insulin Resistance, and Colorectal Cancer: Could miRNA Dysregulation Play a Role?» En: *International Journal of Molecular Sciences* 20.12 (ene. de 2019), pág. 2922. DOI: 10.3390/ijms20122922. URL: http://dx.doi.org/10.3390/ijms20122922.

[288] Sean McNabney y Tara Henagan. «Short Chain Fatty Acids in the Colon and Peripheral Tissues: A Focus on Butyrate, Colon Cancer, Obesity and Insulin Resistance». En: *Nutrients* 9.12 (dic. de 2017), pág. 1348. DOI: 10.3390/nu9121348. URL: http://dx.doi.org/10.3390/nu9121348.

[289] Su Yon Jung y Zuo-Feng Zhang. «The effects of genetic variants related to insulin metabolism pathways and the interactions with lifestyles on colorectal cancer risk». En: *Menopause* 26.7 (jul. de 2019), págs. 771-780. DOI: 10.1097/gme.0000000000001301. URL: http://dx.doi.org/10.1097/GME.0000000000001301.

[290] Caio Mazucanti y col. «Longevity Pathways (mTOR, SIRT, Insulin/IGF-1) as Key Modulatory Targets on Aging and Neurodegeneration». En: *Current Topics in Medicinal Chemistry* 15.21 (ago. de 2015), 2116–2138. DOI: 10.2174/1568026615666150610125715. URL: http://dx.doi.org/10.2174/1568026615666150610125715.

[291] Zhang Wang y Martin Wu. «An integrated phylogenomic approach toward pinpointing the origin of mitochondria». En: *Scientific Reports* 5.1 (ene. de 2015). DOI: 10.1038/srep07949. URL: http://dx.doi.org/10.1038/srep07949.

[292] Keith Baar y col. «Skeletal muscle overexpression of nuclear respiratory factor 1 increases glucose transport capacity». En: *The FASEB Journal* 17.12 (sep. de 2003), págs. 1666-1673. DOI: 10.1096/fj.03-0049com. URL: http://dx.doi.org/10.1096/fj.03-0049com.

[293] Jennifer Permuth-Wey y col. «Inherited Variants in Mitochondrial Biogenesis Genes May Influence Epithelial Ovarian Cancer Risk». En: *Cancer Epidemiology Biomarkers Prevention* 20.6 (mar. de 2011), págs. 1131-1145. DOI: 10.1158/1055-9965.epi-10-1224. URL: http://dx.doi.org/10.1158/1055-9965.EPI-10-1224.

[294] José R. Blesa y col. «NRF-1 is the major transcription factor regulating the expression of the human TOMM34 gene». En: *Biochemistry and Cell Biology* 86.1 (feb. de 2008), págs. 46-56. DOI: 10.1139/o07-151. URL: http://dx.doi.org/10.1139/O07-151.

[295] Katarzyna Skonieczna, Boris A. Malyarchuk y Tomasz Grzybowski. «The landscape of mitochondrial DNA variation in human colorectal cancer on the background of phylogenetic knowledge». En: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1825.2 (abr. de 2012), págs. 153-159. DOI: 10.1016/j.bbcan.2011.11.004. URL: http://dx.doi.org/10.1016/j.bbcan.2011.11.004.

[296] Maria-Jesus Sanchez-Pino. «Mitocondrial dysfunction in human colorectal cancer progression». En: *Frontiers in Bioscience* 12.1 (2007), pág. 1190. DOI: 10.2741/2137. URL: http://dx.doi.org/10.2741/2137.

[297] Jéssica Alonso-Molero y col. «Alterations in PGC1 expression levels are involved in colorectal cancer risk: a qualitative systematic review». En: *BMC Cancer* 17.1 (nov. de 2017). DOI: 10.1186/s12885-017-3725-3. URL: http://dx.doi.org/10.1186/s12885-017-3725-3.

[298] Fidel Díez Díaz y col. «GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines». En: *Mathematics* 9.6 (mar. de 2021), pág. 654. DOI: 10.3390/math9060654. URL: http://dx.doi.org/10.3390/math9060654.

[299] Xiao Bai y col. «Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments». En: *Pattern Recognition* 120 (dic. de 2021), pág. 108102. DOI: 10.1016/j.patcog.2021.108102. URL: http://dx.doi.org/10.1016/j.patcog.2021.108102.

[300] Yong-Yeon Jo y col. «Detection and classification of arrhythmia using an explainable deep learning model». En: *Journal of Electrocardiology* 67 (jul. de 2021), 124–132. DOI: 10.1016/j.jelectrocard.2021.06.006. URL: http://dx.doi.org/10.1016/j.jelectrocard.2021.06.006.

Apéndice A

GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines

Resumen:
Los estudios de asociación del genoma completo (Genome wide-association studies; GWAS) son estudios observacionales de un gran conjunto de variantes genéticas en la muestra de un individuo con el fin de averiguar si alguna de estas variantes está vinculada a un rasgo concreto. En las dos últimas décadas, los GWAS han contribuido a varios nuevos descubrimientos en el campo de la genética. Esta investigación presenta una metodología novedosa a la que pueden aplicarse los GWAS. Se basa principalmente en dos metodologías de aprendizaje automático, algoritmos genéticos y máquinas de vectores soporte. La base de datos empleada para el estudio constaba con información sobre 370.750 polimorfismos de un solo nucleótido pertenecientes a 1076 casos y 973 controles de cáncer colorrectal. Se probaron diez *pathways* con diferentes grados de relación con el rasgo estudiado. Los resultados obtenidos mostraron cómo la metodología propuesta es capaz de detectar *pathways* relevantes para un determinado rasgo: en este caso, el cáncer colorrectal.

Referencia:
F. Díez Díaz, F., F. Sánchez Lasheras, V. Moreno, F. Moratalla-Navarro, A. J. Molina de la Torre, V. Martín Sánchez, **GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines**, Mathematics, vol. 9, pp. 1-19, Mar. 2021, ISSN: 2227-7390, Marzo 2021

DOI:
https://doi.org/10.3390/math9060654