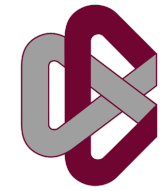


CIMAT

Centro de Investigación en Matemáticas

Métodos de Ciencia de Datos aplicados al diagnóstico del cáncer colo-rectal en población Mexicana

T E S I S

PARA OPTAR POR EL GRADO DE:

Maestro en Cómputo Estadístico

PRESENTA:

Rael Rojas Barrantes

ASESOR:
Dr. José Ulises Márquez Urbina.

CO-ASESOR:
Dr. Augusto Rojas Martínez.

Monterrey, Nuevo León, 2019

TRIBUNAL ASIGNADO

Presidente:	Dra. Graciela María de los Dolores González Farías.
Secretario:	Dr. Rodrigo Macías Páez.
Vocal:	Dr. José Ulises Márquez.

La tesis se realizó en ,CIMAT, Monterrey, Nuevo León.

ASESOR DE TESIS:

Dr. José Ulises Márquez Urbina.

_____

CO-ASESOR DE TESIS:

Dr. Augusto Rojas Martínez.

_____

A mi familia; a mi esposa, quien en todo momento me apoyo y tuvo comprensión y paciencia cuando quise iniciar con esta aventura. En verdad, gracias.

Reconocimientos

A don Joaquin Ortega, quien a pasar de todos los inconvenientes confió en mi, me aconsejo. El mundo seria mejor con muchos profesores, tutores y colegas como él, gracias.

A Jannet Vega y a Ma. Dolores Aguilar de servicios escolares, quienes siempre estuvieron apoyando y respondiendo mis correos cuando tuve dificultades con los procedimientos administrativos, disculpas por las molestias.

A todos mis compañeros y amigos de Guanajuato que aún sin saber como, supieron comprenderme y tenerme paciencia, siempre dispuestos a ayudar a comprender esos puntos que resultaban poco menos que imposibles de asimilar.

A Don Rodrigo, Doña Graciela, Ulises y en general a todo el personal de CIMAT Monterrey, quienes en definitiva son quienes me mostraron la razón de ser de la estadística, su aplicación en ejemplos reales; en definitiva lo que busque desde un comienzo.

A Don Augusto Rojas, Doña Rocio Ortiz, del Hospital San José. Que me brindaron apoyo para iniciar en el tema, me facilitaron los datos y me dieron un primer punto de inicio para aprender, abriendo una nueva puerta de posibilidades. También gracias infinitas a Valentina Colistro de Uruguay, quien estuvo presente aclarándome y escuchando mis puntos de vista, así como dando ideas del como trabajar los datos; del mismo modo a Raquel Cruz de España quien me brindo su opinión y también me oriento sobre que preguntas podrían resultar interesantes de abordar.

A todos mis amigos y compañeros de CIMAT, Monterrey, quienes mostraron siempre que una familia se puede tener fuera de la casa, fuera del país de uno, que me hicieron sentir querido.

Finalmente, al proyecto N 295926 del Laboratorio Nacional de Sistemas Proyecto (para Enfermedades Crónico Degenerativas) y al CONACYT que financiaron mis estudios hasta el final.

Este trabajo es producto de mi interrelación con todas las personas e instituciones anteriores y muchas, muchas más que me hacen ser quien soy. No solo es la culminación de un proceso, sino el inicio de muchos otros.

A todos las gracias.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra institución. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Rael Rojas Barrantes. Monterrey, Nuevo León, 2019

Resumen

El CCR es una enfermedad multifactor con alta mortalidad que la ubican entre la primeras 5 causas de muerte por cáncer en el mundo. Los factores no hereditarios que lo generan son aún desconocidos, sin embargo a lo largo de los años estudios principalmente en Estados Unidos, Europa y Asia, lo han asociado a variantes no genéticas entre ellas: Sedentarismo, el consumo de grasas provenientes de carnes rojas, el alcoholismo, entre otras; estas asociaciones han tenido criterios divididos en la comunidad científica generando expectación sobre el como y en que medida están asociadas.

El presente trabajo explora esas asociaciones mediante el uso de técnicas de Ciencia de Datos. Basado en análisis estadísticos realizados sobre la base de 3525 individuos mexicanos se muestran asociaciones entre las variables genómicas y no genómica mediante un **Análisis de Correspondencia múltiple** (MCA), explorando los beneficios respecto al Análisis por Componentes Principales (PCA); se presenta una variedad de modelos predictivos unos creados mediante regresión logística y aplicando selección de variables mediante LASSO, o técnicas de aprendizaje automático como Random Forest, Adaboost, Soport Vector Machine, entre otros, realizando en cada caso una comparativa entre las ventajas y desventajas de cada uno.

Índice general

Índice de figuras

Índice de cuadros

Introducción

Según la Organización Mundial de la Salud (OMS), el Cáncer Colo-Rectal (CCR) es uno de los cinco de mayor incidencia y mortalidad en el mundo; estudios para determinar sus orígenes como Genome-wide association study (GWAS) y el proyecto de mapeo de haplotipos (Proyecto de Mapeo de Haplotipos (HAPMAP)) han sido llevados acabo con datos de poblaciones europeas, asiáticas y norteamericanas, mostrando evidencias de asociaciones génicas con la patología. Sin embargo es conocido que estas asociaciones varían de país a país y de continente a continente [50], haciendo apremiante la búsqueda de los factores propios de cada región.

El CCR es una enfermedad multifactorial en la que se combinan causas hereditarias con patrones de vida [40, 42, 45, 51]; pero la dificultad para establecer asociaciones directas [33, 38] han generado opiniones divididas respecto al porcentaje de influencia que tienen cada uno de los diferentes factores.

El creciente conocimiento sobre el análisis de datos de gran dimensión, así como la mejora y evolución de los métodos predictivos, amplían la base de estudios tradicionales, para incluir nuevas disciplinas como la informática y la estadística. Estas facilitan el análisis de datos de diferente naturaleza, estableciendo interrelaciones entre las variables y sus efectos. El uso de métodos de aprendizaje automático (Machine Learning) [5, 17, 24] y diferentes técnicas de asociación estadística (como MCA, PCA, CCA, entre otros) permiten establecer relaciones no evidentes que pueden resultar en puntos de partida para análisis posteriores.

El estudio de las causas del CCR y sus efectos tiene varias aristas de importancia; en particular, una corresponde al costo que representa para la seguridad social. Jemal et al. [26] predice 112 800 casos para el 2018 en América Latina, de los cuales 11 376 corresponderían a México. Una política preventiva en torno al CCR podría generar un ahorro (en promedio) de entre $1200 a $92000($USD) por año por persona [2]. Este rango esta dentro del reportado por GNP seguros, entidad que indicó que al 2016 había pagado un total de 110 millones de pesos, por tratamientos ligados a 530 casos de CCR [4].

Los hechos anteriores, aunado a una cultura que se resiste a los métodos de diagnós-

tico tradicionales como el análisis de sangre oculta en heces y la colonoscopia [15, 35, 45], junto con una aparición cada vez más temprana del CCR, hacen apremiante establecer nuevas pautas que permitan ir un paso adelante en la búsqueda de los mecanismos de combate.

Con el fin de determinar causas genéticas y no genéticas asociadas al CCR para la población mexicana, el presente trabajo analiza dos bases de datos originadas de una misma muestra y resultado del proyecto Genetic study of Common Hereditary Bowel Cancers in Hispania and the Americas (CHIBCHA), financiado por la Unión Europea. Se analizan los datos de 3525 pacientes mexicanos, de los cuales 1266 estaban diagnosticados con CCR. A partir de está información, en esta tesina se pretende responder a las preguntas:

1. *¿Es posible determinar asociaciones genéticas y ambientales en las bases de datos mexicanas que permita diagnosticar la susceptibilidad al padecimiento de CCR?*
2. *¿Cómo están relacionadas las variables genéticas y ambientales con su efecto al CCR?*
3. *¿La población mexicana analizada coincide con los patrones internacionales de tenencia de CCR?*

Con el fin de responder lo anterior, se plantea como objetivo de investigación: establecer asociaciones entre SNPs analizados y las variables no genéticas con el Cáncer Colo-Rectal en población mexicana.

Para ello se desarrollan cada uno de los siguientes pasos o fases:

- Análisis exploratorio de las bases de datos presentes.
- Determinación de grupos de variables con marcada importancia en la diferenciación del CCR.
- Estudio de las interrelaciones de los predictores, así como su influencia con el fenotipo.
- Generación de un perfil genético y clínico para la población mexicana, a partir de un modelo matemático, que describa el comportamiento de la patología.
- Validación de los resultados internacionales para el establecimiento de las políticas de salud.

Los Capítulos 1 y 2 se centran en la descripción de los esfuerzos realizados alrededor del problema y similares; así como a justificaciones teóricas de los métodos utilizados.

En el Capítulo 3 se realiza una descripción de cada una de las bases suministradas, así como el proceso de limpieza y formato de las mismas. Se describen los métodos de

imputación utilizados para cada base, así como la agrupación de variables.

El Capítulo 4 establece asociaciones entre las variables mediante un Análisis de Correspondencia Múltiple (MCA); se determinan grupos de variables con efecto protector, así como posibles variables asociadas a un efecto causal.

En el Capítulo 5 se amplía sobre las relaciones entre los predictores, se establecen porcentajes de participación en la predicción por parte de cada conjunto de datos y se hace uso de diferentes métodos de aprendizaje de máquina (Machine - Learning) generando modelos de clasificación para la patología, comparando su eficiencia mediante el uso de curvas de recepción operativa de características (ROC).

Finalmente en el Capítulo 6 se realiza una recopilación de los resultados y se indican dificultades y limitaciones encontradas. Se termina dando recomendaciones sobre posibles estudios posteriores en torno al tema.

Capítulo 1

Estudios asociados a la determinación de patologías

En esta sección se realiza un breve repaso de algunos de los conceptos genéticos utilizados, con el fin de poner en contexto el proceso científico seguido en el estudio de los nucleótidos, se pretende establecer una base sobre su relevancia en el desarrollo de diversas enfermedades.

Se abordan diferentes estudios respecto al análisis del genoma y sus asociaciones con diferentes enfermedades, con énfasis en el estudio del cáncer y las asociaciones con variables genéticas y no genéticas.

1.1. Conceptos preliminares

Los nucleótidos, como bloques de "lego" corresponden a constituyentes básicos de todo ser vivo, estos o su combinación modelan las funciones de un organismo pudiendo ser los causantes de patologías genéticas. La generación de mutaciones producto de la exposición a agentes externos puede conllevar al desarrollo de enfermedades, las cuáles si bien son generadas mediante la afectación de componentes genéticos no son asociadas a estos, sino a sus causantes ambientales. La determinación de estas causantes, sus afectaciones y consecuencias han sido estudiadas mediante diversas técnicas, principalmente a partir de la secuenciación del genoma humano.

Un *gen* por su parte, corresponde a una sección del ADN el cual a su vez corresponde a un conjunto de nucleótidos conformados por una molécula de azúcar (desoxirribosa), un compuesto de fósforo (grupo fosfato) y una base nitrogenada (adenina A, citosina C, guanina G y timina T). La Figura 1.1 muestra estas relaciones.

Estos compuestos se organizan de forma tal que la desoxirribosa y el grupo fosfato

conforman una escalera en forma de doble hélice, en la cuál los peldaños o genotipos corresponden a las combinaciones A-T, C-G en este contexto llamadas "bases" o par de bases. El conjunto de todas estas combinaciones para una especie se denomina secuenciación genética, La secuenciación de solo algunas regiones de interés - llamadas *loci*- se denomina *genotipado*.

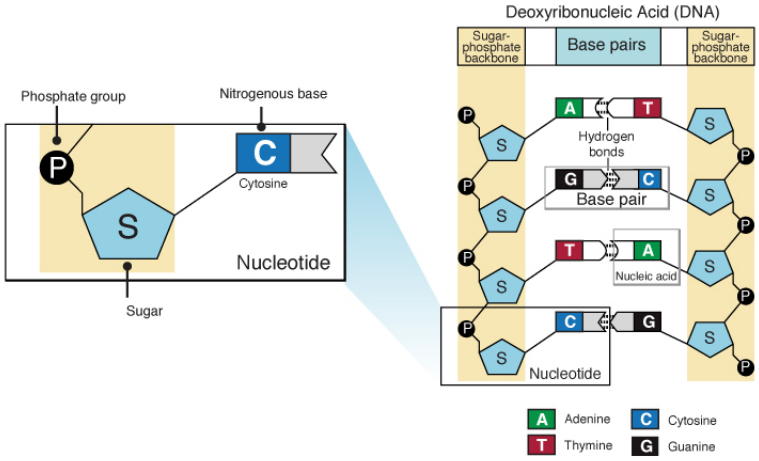


Figura 1.1: Conformación de un nucleótido. El conjunto de cientos o miles de pares de bases (pb) conforman a los diferentes genes. Tomado de www.genome.gov

Cuando se genera una modificación de alguna base(locus) afectando al menos el 1 % de la población en estudio, se dice que se ha generado un polimorfismo. Estos son de interés dado que su modificación afecta la constitución de aminoácidos, los cuales a su vez afectan el funcionamiento de las proteínas que constituyen al gen.

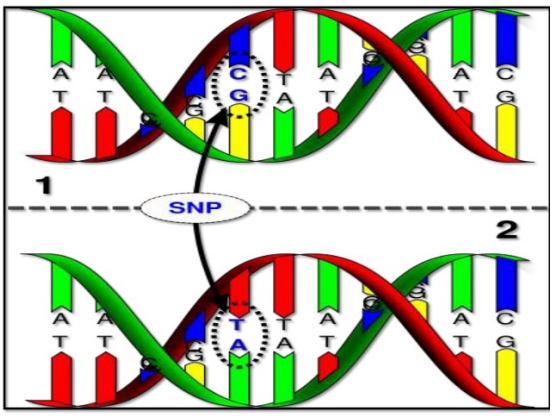


Figura 1.2: Si las variaciones alcanzan a más del 1 % de la población se considera un polimorfismo de nucleótido simple (SNP). Imagen: Bajo derechos de código libre CC

Durante el proceso de transcripción del ADN la estructura de los alelos es conservada casi sin variación producto de los procesos regulatorios, de reparación y edición molecular, en este sentido, cuando se genera un polimorfismo en la cadena y los genes encargados de su corrección se encuentran afectados, estas se transmiten a lo largo de las próximas generaciones dado origen a la variabilidad genética. La modificación generada puede ser de sentido erróneo, sin sentido o de inserción de bases. La Figura 1.2 muestra una mutación de sentido erróneo donde la base C-G ha sido modificada por T-A, lo cual genera la modificación de aminoácidos. Estas variaciones entre individuos estudiadas a partir de un conjunto de locus de un mismo alelo es conocida como haplotipos.

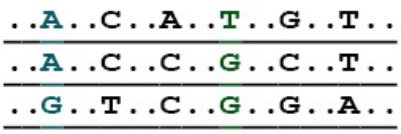


Figura 1.3: La primera y cuarta letra de cada fila (A-T, A-G, G-G) son suficientes para identificar cada uno de los tres haplotipos. Tomada de www.genome.gov

La distribución de los haplotipos en el genoma permite identificar orígenes poblacionales como en el caso de la Figura 1.3 la cual se muestra 3 haplotipos cuyas variantes entre bases los diferencian.

1.2. Proyecto Genoma Humano (PGH)

En el 2003 se completó la primera versión de (PGH) el cual determinó que el ADN del ser humano consta de aproximadamente 3000 millones de bases (Mb) en las cuales los genes están distribuidos en promedio cada 100 mil bases (kb). Cada gen tiene una longitud media que varia entre 20 kb y 30 kb y actualmente se conoce la función aproximadamente de un 5 % de ellos [3].

El PGH es uno de los proyectos que determina que los seres humanos son idénticos en 99 % de su material genético, mientras que un 1 % es el que genera la variabilidad fenotipica (correspondiente a los rasgos que se manifiestan o son observables). Esta variabilidad depende de la existencia de aproximadamente 10 millones de SNP y sus interrelaciones [36], las cuales pueden generan disposición o resistencia ante algún tipo de padecimiento.

Si esta variación es heredable y genera una modificación observable entonces se denomina mutación o mutación génica, en el caso de que la afectación sea sobre un gen en especifico. Las mutaciones génicas asociadas a enfermedades de forma directa

se les conoce como *mutaciones de alta penetrancia*, las cuales en el caso del CCR sólo permiten explicar el 1 % de los casos diagnosticados [24].

1.2.1. Proyecto de Mapeo de haplotipos (HapMap)

El Proyecto HAPMAP iniciado en el 2002, tenía como objetivo crear un mapa de haplotipos del genoma humano, describiendo los patrones comunes de variación, las regiones cromosómicas con conjuntos de SNP fuertemente asociados, los haplotipos en esas regiones y los SNP que marcan esos haplotipos. Finalizado en el 2005, cumplió sus objetivos agrupando a diferentes centros de investigación de Canadá, Estados Unidos, Reino Unido, Nigeria, Japón y China; proporcionando un recurso que permite determinar los genes que afectan a la salud.

Mediante la comparación de poblaciones con las presentes en la base del proyecto, se logra identificar entre 250K a 500K SNP. Esta cantidad aporta casi tanta información como la obtenida a realizar un estudio de secuenciación a los 10 millones de SNP de un individuo.

Al respecto, un estudio de Desequilibrio de Ligamiento (LD) consiste en analizar el grado de variación de un conjunto de nucleótidos de un mismo cromosoma, si estos presentan asociación entonces se indica que la presencia de uno de ellos esta ligada con la presencia del otro, es decir, existe un **"bloque" que se transmite de generación en generación con poca variación**. Con esta información se puede caracterizar un haplotipo -el cual correspondería a un bloque o conjunto de nucleótidos- a partir de la presencia de los SNP que presenten un alto Desequilibrio de Ligamento (DL).

1.2.2. Estudio de Asociación de Genoma Completo (GWAS)

En busca de las relaciones de los polimorfismos de nucleótido simple (SNP) con enfermedades, los GWAS analizan todo el genoma mediante el uso de grandes bases de secuenciación [34, 36] -100mil a 200mil sujetos o más- que permitan determinar diferencias significativas en desordenes genéticos poco frecuentes, como diferentes tipos de cáncer, enfermedades mentales, entre otros. El uso de gráficos tipo Manhattan es frecuentemente utilizado. En la Figura 1.4, se observan valores particularmente fuertes a partir de p-valores en el eje Y para el cromosoma 10 en un estudio de asociación para cáncer de seno en población japonesa.

El gran número de muestras y la necesidad de un robusto sistema que evité o en su defecto minimice los falsos positivos, hacen costosos los estudios de asociación genética. A pesar de estos limites, los GWAS han permitido determinar *loci* que presentan alta variabilidad génica logrando asociarlos con enfermedades y facilitando su comprensión [36].

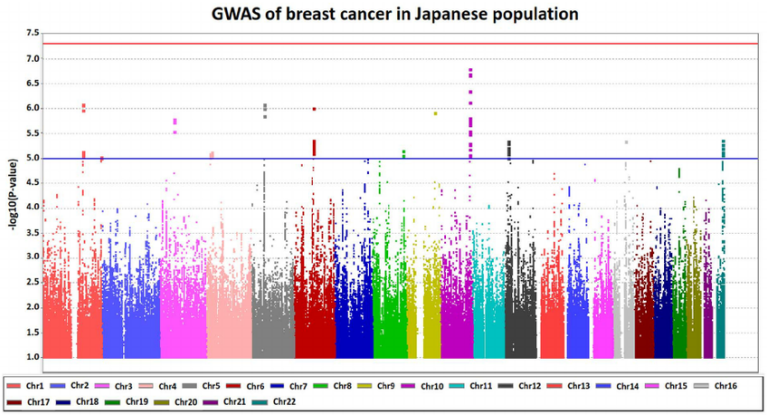


Figura 1.4: Gráfico Manhattan de un estudio de Asociación para pacientes de Cáncer de seno en Japón [34]. La altura de los puntos es proporcional a la significancia del SNP.

Los GWAS han sido aplicados principalmente en poblaciones europeas, asiáticas y norteamericanas, permitiendo explicar que entre 5 % y 10 % de los casos de CCR están asociados a causas hereditarias como la políposis adenomatosa familiar y el síndrome de Lynch [38] .

Lichtenstein et al. [33] por su parte, establecen que se puede asociar hasta un 35 % de los casos de CCR a variables genéticos-hereditarias, abriendo la posibilidad de existencia de interacciones genéticas y ambientales ocultas. Estas interacciones entre genes mutados que no inciden directamente en el desarrollo de una enfermedad sino que se presentan producto de sus interacciones se denominan *mutaciones de baja penetrancia*. La repetición de estos genes en personas diagnosticadas podría dar indicio de la susceptibilidad al padecimiento.

1.2.3. Proyectos 1000 genomas y 100K genomas

El proyecto 1000 genomas -iniciado en 2008- amplia el panorama desarrollado por HAPMAP al incluirse el mapeo genético de poblaciones de diversas latitudes - específicamente 26, Figura 1.5 -, cubriendo el 99 % de las variantes haplotipicas [7] que permiten dilucidar parte de las lagunas existentes en torno al efecto de los genes de baja penetrancia y su peso conjunto en la aparición de una patología. El proyecto contó con la participación de 2504 individuos secuenciados en grupos de 50 pares de bases.

Figura 1.5: Poblaciones consideradas en 1000 Genomas. Tomado de IGSR.

De manera similar, 100k genomas corresponde a un proyecto piloto de Reino Unido que pretende determinar la secuenciación completa de 100 mil pacientes de los centros de salud. El proyecto se centra principalmente en enfermedades raras y cáncer, las cuales afectan a entre un 6 % y 7 % de la población.

Los proyectos 1000 y 100k son de suma importancia dado que la comparativa de sus resultados permite obtener información sobre lo que hace diferente a cada individuo y el cómo sus variantes genéticas, producto de la descendencia o de su estilo de vida, lo afectan. Es precisamente el hecho de que seamos diferentes lo que dificulta discernir entre las variables que generan susceptibilidad a una enfermedad. En este sentido, entre más definidos (o detallados) estén los mapas genéticos de una población en particular, más probable será el determinar las variables de interés para esa población.

1.3. Antecedentes y estudios sobre el CCR

Existen diferentes interpretaciones de lo que es el cáncer y su origen, la forma más habitual de referirse al mismo corresponde a su caracterización. En particular, The American Cancer Society (ACS) [44] lo describe como *"[Un] crecimiento descontrolado de células anormales"* . De forma similar, la Organización Mundial de la Salud (OMS) indica

> *El cáncer implica una ruptura patológica de los procesos que controlan la proliferación celular, la diferenciación y la muerte de células.(...) las células malignas que forman un tumor surgen de tejido epitelial (...) y se denominan "carcinoma."* Stewart et al. [45, p.12]

De lo anterior se infiere que, el cáncer refiere a una modificación en la naturaleza de los procesos celulares, misma que se puede originar por diferentes situaciones como las presentadas en el Cuadro 1.1.

The American Cancer Society	**Organización Mundial de la Salud**	
▪ Tabaco	▪ Tabaco	▪ Radiación, alimentación
▪ Organismos infecciosos	▪ Consumo de alcohol	▪ Infecciones crónicas,
▪ Alimentación no saludable	▪ Exposición ocupacional	▪ Causas por inmunosupresores
▪ Mutaciones genéticas heredadas	▪ Contaminación en el aire	▪ Susceptibilidad genética, factores reproductivos y hormonales.
▪ Hormonas y afecciones inmunitarias.	▪ Agro-químicos, fármacos	

Cuadro 1.1: Causas del cáncer Colo-Rectal por organización.

Las causas presentadas pueden actuar en conjunto o de forma aislada, provocando diferentes respuestas en el organismo. En el caso especifico del Cáncer Colo-rectal (CCR), la asociación entre las anteriores no es del todo clara [27] generando controversia en su aporte.

El CCR es un problema mundial que afecta anualmente a aproximadamente un millón de personas con una mortalidad de aproximadamente el 50 % [50]. El riesgo varia por país y por continente, los reportes actuales parecen coincidir en que factores hereditarios, dieta y el estilo de vida están asociados.

En la carciogénesis del CCR se destaca principalmente la aparición de pólipos adenomatosos. Sin embargo, fuera de las causas hereditarias -las cuales contribuyen en un porcentaje relativamente bajo- , su origen es parcialmente incierto; resultando, en contraste, que la contribución de las variables genéticas y sus relaciones con los factores no genéticos pueden ser una vía de respuesta, de la cual se tiene conocimiento limitado [33].

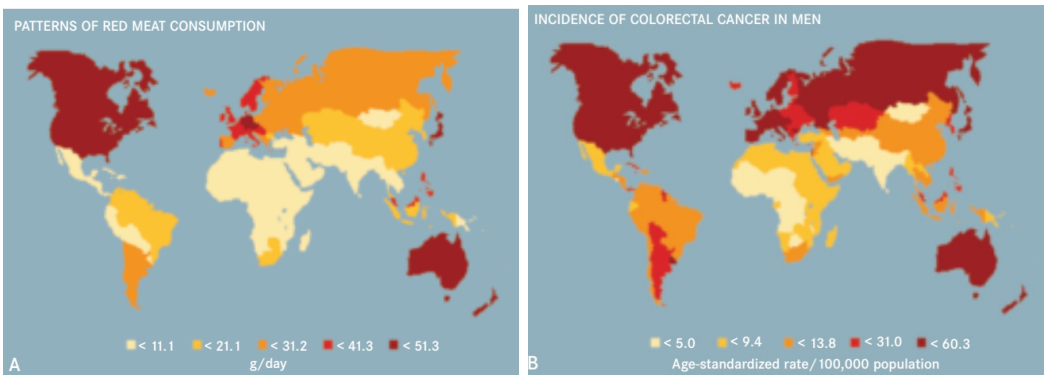
1.3.1. Variables no genéticas

Se considerará variables no genéticas a aquellas modificaciones que afectan a los gene pero que no corresponden a un efecto hereditario, como lo son la alimentación, el ejercicio físico, los hábitos de fumado, el consumo de alcohol, la edad, el género. Cabe la pregunta: ¿cómo se relacionan este tipo de variables con el CCR?

Actualmente, se estima que alrededor del 90 % de los casos de CCR ocurren a partir del desarrollo de *adenocarcinomas* de forma esporádica [ver 45, p198] , de los cuales un 80 % son de tipo lieberkühniano (originados en la cripta del colon) y un 10 % de tipo coloide mucoso [38]. Se estima que de estos diagnosticados entre un 60 % a 80 % se pueden prevenir adoptando hábitos de vida saludables [27, 44, 45]. A continuación se detallan algunas de estas influencias.

Alimentación

De acuerdo con Stewart et al. [45, pág 62] existen indicios de que aproximadamente el 30 % de los casos de cáncer diagnosticados están asociados a una dieta pobre en frutas y vegetales. El consumo diario de 500g de estás, puede reducir la incidencia en alrededor de un 25 % de los casos asociados a los tipos de cáncer del tracto digestivo.



(a) ***Consumo de Carne Roja*** **(b)** ***Incidencia del CCR***

Figura 1.6: Al realizar las comparativas entre el consumo de carne y la incidencia del CCR, se observa un patrón de correlación. Sin embargo estas relaciones no son concluyentes en todas las poblaciones lo que genera controversia en los resultados. Tomado de [45, pág 65]

Particularmente para el consumo de carne roja y procesada, Stewart et al. indican que se ha determinado una incidencia directa con el desarrollo del CCR a partir del consumo 80g de carne al día, el cual puede incrementar el riesgo entre un 25 % a un 67 %. En su guía práctica para el diagnóstico del CCR, la Secretaría de Salud mexicana

[42] va mas allá, indicando que no es solo el consumo de carne roja, sino aquella que tiene alto contenido de grasa animal.

La Figura 1.6 muestra esta relaciones para la Organización Mundial de la Salud. Sin embargo, este hecho todavía genera controversia [27] pues se obtienen resultados no concluyentes en diferentes poblaciones. El considerar que la asociación del CCR con la alimentación no sea inmediata en todas las poblaciones, hace pensar la existencia de variables ocultas las cuales pudieran estar generando mecanismos protectores o de susceptibilidad asociados a la alimentación.

Obesidad y sedentarismo

En un estudio longitudinal (1962 - 1988) realizado a 17 595 alumnos de Harvard por Johnson y Lund [27], en el cual se estudia la incidencia de la nutrición y los estilos de vida con el CCR, Se indica que el riesgo de padecer CCR asociado al quintil de mayor peso fue casi 2.5 veces al riesgo presentado en el quintil más ligero. Sin embargo, los autores argumentan que el peso por si sólo no muestra evidencia en el riesgo de padecer CCR, sino que la medición debe realizarse en conjunto con la práctica de actividad física; al respecto se menciona que el solo hecho de realizar actividad física moderada de al menos 2 horas diarias - o 1 hora de actividad intensa- muestra una reducción de alrededor de 20 % en la susceptibilidad de padecer CCR, excepto en los casos en que los individuos fueran mujeres sometidas a terapia hormonal.

Uno de los resultados más interesantes del estudio presentado por Johnson y Lund corresponde a observar niveles anormalmente altos de *leptina* en las personas con obesidad. La leptina es una proteína relacionada con el proceso de muerte celular (apoptosis) generada por las células adiposas, lo que podría conllevar a establecer relaciones entre la obesidad y la presencia de células cancerígenas.

Consumo de alcohol y tabaco

A nivel mundial se ha asociado el consumo de alcohol al cáncer de boca, esófago e hígado y se ha establecido evidencia con el cáncer de seno y CCR. Para todas las asociaciones de cáncer con el consumo de alcohol se ha establecido que el riesgo crece linealmente con el consumo a partir de 80 ml por día (en el caso del vino a partir de 1 litro).

Por otro lado, el fumado es conocido como causante de muertes en el mundo, particularmente a partir de cáncer de pulmón [45, p23], en el cual el riesgo de contraer cáncer es proporcional al tiempo de fumado. Para una persona que fuma entre 21 a 39 cigarrillos por día e inicia a los 15 años hay un 50 % más de probabilidad de contraer cáncer que una persona inicia a los 25 años. En el caso particular del CCR, el tabaquismo no

es una causa directa; sin embargo su asociación como factor de riesgo en el desarrollo de pólipos adenomatosos [42] lo hacen ser considerado en los estudios.

1.3.2. Variables genómicas

Estudios internacionales indican que hasta un 35 % de los casos de CCR corresponden a causas genético-hereditarias [33, 40]. Sin embargo, las causas hereditarias conocidas corresponden a un porcentaje bajo del total de casos [45, 50] -de un 1 % a un 5 %-.

De acuerdo con Rúa [40], se observa que la gran heterogeneidad genética del CCR conlleva a considerar varias vías en su desarrollo, las cuales están principalmente asociadas a genes reparadores, supresores o oncogenes -genes multifunción los cuales son altamente mutables- . Algunas de las principales vías [51] corresponden a:

1. Vía supresora (inestabilidad cromosómica): Ocurre en alrededor del 70 % de los casos de CCR esporádico; consiste en una modificación en el número de cromosomas o afectación estructural en los cromosomas (deleciones o translocaciones, entre otras). Tradicionalmente esta asociada a mutaciones en el gen TP53 ,APC o K-Ras.

2. Vía mutadora (inestabilidad microsatelital): Presente en cerca del 20 % de los casos de CCR, corresponde a un cambio en el número de secuencias cortas del ADN (llamados microsatélites) generando fallos en los sistemas de reparación normal del ADN. La inestabilidad microsatelital presentada en algunos genes como: TGF- βR2, Bax, Caspase 5, MSH3, MSH6, β-catenina, APC, IGFII, y E2F4 están fuertemente asociado a CCR.

3. Vía metilación: Estudiado en cerca del 15 % de los casos de CCR [14], consiste en la unión de un grupo metilo en el carbono 5 de la citosina por acción de las encimas ADN-metiltransferasa. Las modificaciones por metilación generan la inactivación en la expresión de diferentes genes de importancia como el MLH1 (asociado a modificaciones propias de inestabilidad microsatelital hereditarias). La metilación de genes como MINT6, MINT24, MINT32, ER esta asociado a un incremento de la edad.

Al respecto el modelo genético de Vogelstein [14], mostrado en la Figura 1.7, describe varias vías genéticas de la carciogénesis. En particular la vía supresora -señalada con una flecha- establece al menos 7 etapas en el desarrollo del CCR:

- **Etapas 1-2:** A partir de un epitelio normal se genera una mutación (la cual puede estar ligada al historial familiar del gen APC o de forma esporádica) generando una hipermetilación del epitelio, lo que conduce a la inactivación de diversos genes (reparadores o reguladores principalmente) generando un adenoma temprano.

- **Etapas 3-4:** Se genera una modificación en el funcionamiento del oncogen K-ras (en un 75 % de los casos a perdida del cromosoma 17p) lo que conduce a un adenoma intermedio. 50 % de los casos de CCR están asociados a este gen.
- **Etapa 5:** Se genera pérdida de la función de genes supresores ubicados principalmente en el cromosoma 18q generando un adenoma tardío.
- **Etapa 6:** Se genera pérdida de la función de genes supresores ubicados principalmente en cromosoma 17p generando el carcinoma.
- **Etapa 7:** Se generan otras alteraciones genéticas que llevan a la metástasis.

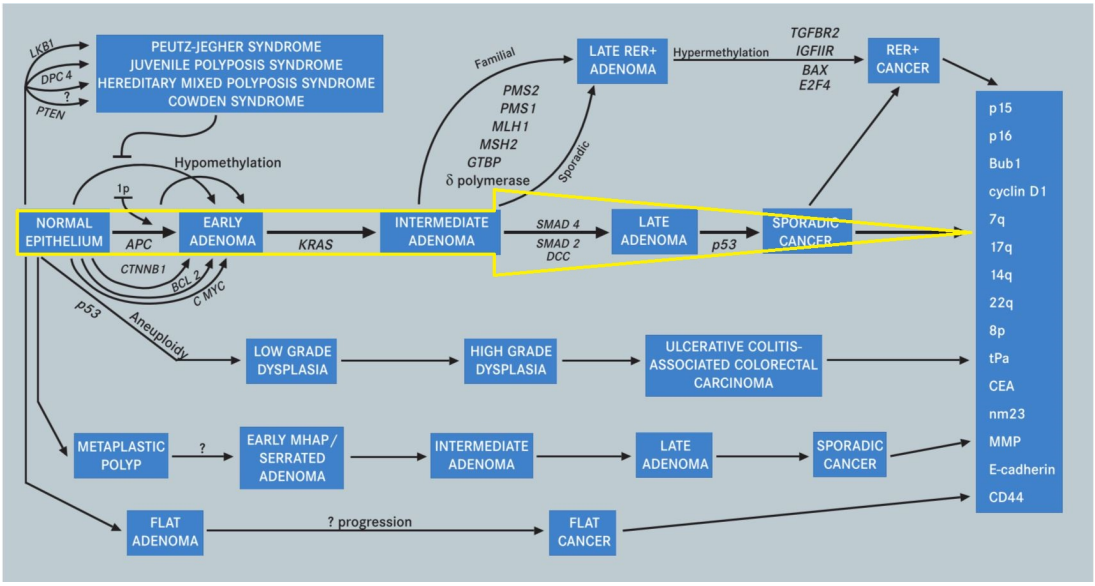


Figura 1.7: Modelo de mutaciones secuenciales acumulativas de Vogelstein. Tomada de [45]

Como indica Worthley et al. [51], los casos de CCR se originan principalmente por la vía de inestabilidad cromosómica, en la cual están implicados principalmente los genes APC (5q21), K-RAS (12p12), entre otros. Sin embargo existen otras asociaciones de interés, en particular la Figura 1.2 presenta una breve descripción de algunos de los genes implicados en el modelo de Vogelstein que han mostrado asociación con el CCR.

Gen	Participación	Funciones Asociadas	Comentarios
ACP	30 % - 85 %	▪ Migración y adhesión celular ▪ Segregación cromosómica ▪ Control del ciclo celular ▪ Reparación del ADN.	Es encargado de prevenir la acumulación del β - catenina cuando la vía Wnt se encuentra inactiva, su mutación genera pérdida de heterocigosidad, es de tipo tipo Supresor - dominante y se ubica en 5q21. Mutaciones en el exón 15 (representan el 75 %) conducen a la inactivación del gen.
K-RAS	30 %-50 %	▪ Transducción de señales del interior al exterior de la célula. ▪ Actúa como interruptor celular que modifica la expresión de múltiples genes reguladores. ▪ Inducción apoptosis. ▪ La quimiotaxis.	Mutaciones puntuales ocurren en el 98 % de los casos en los codones 12 y 13 asociados a una resistencia en las terapias anti-EGFR, es un gen de tipo oncogen y se ubica 12p12.
TP53	25 % - 75 %	▪ Parada de ciclo celular. ▪ Inducción de apoptosis. ▪ Reparador del ADN. ▪ También participa en la transcripción de genes.	Su afectación genera mutaciones germinales, es de tipo supresor y se ubica en 17p13, se encuentra asociado a cerca de un 50 % de varios tipos de cáncer, entre ellos el CCR.
DCC	$\sim$ 60 %	▪ Involucrado en adhesión celular. ▪ Inducción apoptosis.	Problemas con este gen provocan pérdida de heterocigosidad. Es de tipo recesivo supresor, se ubica 18q21. Sus mutaciones se deben principalmente a la pérdida alélica.
SMAD 2-4	$\sim$ 60 %	▪ Codifican proteínas unión-receptoras (SMAD-2 y SMAD-3). ▪ SMAD-4 Participa en conjunto con las R-SMAD permitiendo la translocación del núcleo. ▪ Importantes en la regulación de la apoptosis.	Alteraciones en SMAD-4 se presentan en aproximadamente 20 % de los casos de CCR, generando polípo-sis juvenil cuando ocurren en linea germinal. SMAD2 y 4 se presentan en 18q21, mientras que SMAD3 en 15q22.33.

Cuadro 1.2: Principales genes asociados al CCR. Adaptado de [40]

1.3.3. Estudios latinoamericanos

Los estudios genéticos relativos a enfermedades y sus asociaciones han estado ligados tradicionalmente a poblaciones europeas, asiáticas o estadounidenses. En el caso latinoamericano, y en particular el mexicano, los estudios y proyectos referentes al análisis genómico y asociación genética son recientes o se encuentran en fase de investigación.

El proyecto CHIBCHA, financiado por la Comunidad Europea y finalizado en 2013, es uno de estos. A pesar de haber concluido, algunos grupos de investigación en distintos de los países que conformaban el consorcio (México entre ellos) aún se encuentran en el análisis de los resultados. El proyecto tenía como uno de sus objetivos determinar asociaciones genéticas con diferentes tipos de enfermedades, así como servir de recurso de capacitación para futuros profesionales en el área (según se indica en el sitio del proyecto `cordis.europa.eu`). Particularmente en el caso del CCR, este estudio ha determinado las asociaciones con los *loci* **8q23.3, 8q24.21, 10p14, 11q23.1, 15q14 y 18q21**, ayudando a aclarar la pregunta ¿Dónde buscar?

Otros estudios encontrados son relativos al desarrollo y evolución del CCR, así como su tratamiento [21, 35] en cuyo caso el diagnóstico preventivo se ve limitado al uso de recomendaciones internacionales, existiendo controversia en la frecuencia y aplicabilidad de algunas de las técnicas [42].

El presente trabajo se basa en CHIBCHA y cuenta con el apoyo del Laboratorio Nacional de Medicina de Sistemas (para Enfermedades Crónico Degenerativas) y de distintos de los grupos de investigación en particular en Uruguay y España.

Capítulo 2

Técnicas de modelación y predicción

En esta sección se da una breve introducción a los métodos de ciencia de datos utilizados para la creación y escogencia de los modelos de clasificación, así como el contexto de su utilización y los métodos de evaluación utilizados en los capítulos 4 y 5.

Se abordan puntos concernientes a las interpretaciones dadas a los resultados de los siguientes capítulos, así como justificaciones sobre la selección de parámetros. Información relativa a este capítulo se puede consultar en [1, 9, 11, 12, 18, 25], así como en las referencias adicionales señaladas al interior del texto.

2.1. Modelo logístico

Considere un modelo en el cual la variable dependiente $\mathbf{Y}$ sigue una distribución *Binomial* $y_i \sim B(p_i, n_i)$ En donde el parámetro $\mathbf{p_i}$ representa la probabilidad de *éxito* $\mathbf{P}(y_i = 1)$ de $\mathbf{n}$ ensayos *Bernoulli* independientes. Note que $0 \leq p_i \leq 1$ al ser una probabilidad, luego la trasformación $\ln\left(\frac{p_i}{1-p_i}\right)$ permite establecer una relación lineal al ampliar el rango de valores que puede tomar

$$logit(p_i) = \ln(\frac{p_i}{1-p_i}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_i \tag{2.1}$$

en donde cada x_i corresponde a las n variables predictoras, p_i a la probabilidad de presencia o ausencia de la combinación de variables observadas y β_j a los pesos o importancia de cada uno de los x_i. Los coeficientes de este modelo se determinan mediante el método de mínimos cuadrados, mientras que la probabilidad p_i usualmente mediante máxima verosimilitud y esta dada por:

$$p_i = \frac{\exp\left(\beta_0 + \sum_{i=j}^{n} \beta_j x_i\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{n} \beta_j x_i\right)} = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^{n} \beta_j x_i\right)} \tag{2.2}$$

En el caso de datos categóricos -como el que nos ocupa- el modelo logístico es aplicable al realizar la codificación de las variables predictivas como variables tipo *dummy* [ver 1, cap 5].

2.1.1. Ajuste del modelo

El problema de sobreajuste aumenta el sesgo y reduce el poder predictivo de un modelo, del mismo modo la falta de ajuste genera un aumento en la varianza. Escoger cuales variables son explicativas y cuales generan un efecto confusor es fundamental para evitar este tipo de problemas. En particular para un modelo de clasificación logística los métodos de Elastic Net o la agrupación de variables mediante mínimos cuadrados parciales (PLS) son algunos de los propuestos en la presente sección.

Elastic Net

Elastic net corresponden a una combinación entre la penalización cuadrática -el método de Ridge - y la penalización lineal -método LASSO - en la cual se busca determinar los valores de β que minimizan la función de error $f(\alpha, \beta)$

$$f(\alpha, \beta) = ||y - X\beta||^2 - (1-\alpha)\sum \lambda_1 ||\beta||^2 - \alpha \sum \lambda_2 |\beta| \tag{2.3}$$

Es importante destacar que el método de Ridge ($\alpha = 0$) no reduce los coeficientes a cero (en algunos casos hecho deseable), caso contrario con Lasso. Los parámetros α y λ se optimizan con el fin de minimizar f [18, pag, 61].

Mínimos cuadrados parciales (PLS)

El método PLS es utilizado para reducir la dimensionalidad de los datos y determinar variables latentes, similar al método de componentes principales o al análisis factorial, es utilizado principalmente cuando se tiene un mayor número de variables predictivas que observaciones.

PLS, a diferencia de PCA, incluye a la variable respuesta dentro del análisis lo que le permite determinar asociaciones directas. En el caso de modelos de clasificación se tienen a utilizar en combinación con otra técnicas como Análisis de Discriminante (DA).

Considere una matriz de predictores $\mathbf{X}$ y una matriz de variables dependientes $\mathbf{Y}$. Suponga que para $\mathbf{X}$ e $\mathbf{Y}$ existe un conjunto de variables latentes $\mathbf{T}$ tal que

$$\begin{aligned} Y &= T\beta_1^t + \epsilon_1 \\ X &= T\beta_2^t + \epsilon_2 \end{aligned} \tag{2.4}$$

donde $\beta_{\mathbf{i}}$ corresponde a las matrices de carga y $\epsilon_{\mathbf{i}}$ a las matrices de errores. La matriz $\mathbf{T}$ se define de forma tal que $\mathbf{T} = \mathbf{X}\beta_{\mathbf{3}}$ con $\beta_{\mathbf{3}}$ la matriz de vectores que maximizan a la matriz de covarianza $\mathbf{X^t} \cdot \mathbf{Y} \cdot \mathbf{Y^t} \cdot \mathbf{X}$

De la matriz $\beta_{\mathbf{3}}$ se aplica una rotación tipo *varimax* en la cual se preservan las variables más significativas para el número de dimensiones seleccionadas de la matriz de covarianza.

2.2. Modelos de Ensemble

Corresponden a un conjunto de métodos cuyo fin es aumentar la eficacia de clasificación a partir de la unión de un conjunto de métodos, mejorando no solo la exactitud, sino que también reduciendo la varianza y el sesgo. Estos métodos evitan el sobre ajuste de los clasificadores a partir de la combinación de varios modelos [9] que utilizan un mismo conjunto de datos para ser creados y que tienen un mismo objetivo.

Se podría interpretar que estos métodos intentan determinar una función objetivo $\mathbf{G}(\mathbf{x})$ que mantenga un balance entre la función de regularización $\mathbf{R}(\mathbf{x})$ y la función de pérdida $\mathbf{L}(\mathbf{x})$

$$G(x) = L(x) + R(x) \tag{2.5}$$

La función de regularización controla el número de divisiones realizadas en cada árbol -modelo base que tiene a utilizarse- reduciendo la varianza, mientras que la función de pérdida mejora el ajuste mediante los datos de entrenamiento.

A partir de la combinación de uno o varios métodos de clasificación utilizados como base y de la selección de los conjuntos de entrenamiento, se selecciona solo algunas de la variables para crear un clasificador débil (no robusto) $\mathbf{G_i(x)}$, al repetir reiteradamente el proceso y utilizando la función de penalización en cada modelo creado para determinar una ponderación de la importancia del clasificador creado $\alpha_{\mathbf{m}}$, se logra obtener un clasificador robusto.

$$G(x) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{2.6}$$

El balance entre la función de regularización y la función de pérdida depende de la selección de las observaciones, así como del factor de ponderación utilizado. Dos orientaciones básicas en los métodos de ensemble son **el bagging y el boosting** [1], a saber:

- Los métodos que utilizan **bagging** se basan en seleccionar al azar del conjunto de datos un subconjunto de entrenamiento -el cual admite reemplazo- para la creación

[1] [12, cap 9] hace una descripción un poco más detallada de cada uno de ellos

de cada modelo el cual es independiente de los siguientes modelos creados, esto evita el sobre ajuste a partir de la independencia de los modelos.

- Los métodos que utilizan **boosting** también utilizan un subconjunto de entrenamiento para cada modelo -sin reemplazo -, pero cada modelo resultante se pondera (se le agrega un coeficiente) para reducir el error; es decir, los modelos son dependientes entre sí, lo que disminuye el sesgo.

En general una de las principales diferencias de cada enfoque (Bagging - Boosting) corresponde a la forma en que atacan el problema de ajuste (Reducción del sesgo o reducción de la varianza) Como menciona Shirai et al. [43]: *Baggins es efectivo para algoritmos con bajo sesgo y alta varianza, mientras que el boosting es efectivo para algoritmos con baja varianza y alto sesgo.*

Existen varios métodos que siguen las ideas anteriores: Random Forest (RF), Adaboost (ADA), Xtreme Gradient Boosting (XGB) son algunos de los más utilizados, los cuales se basan principalmente en árboles de decisión.

- **Random Forest:** Corresponde a un método de bagging, en el cual se entrenan de forma independiente arboles de decisión con un número de predictores pre establecido en cada árbol y escogidos de forma aleatoria -típicamente $\sqrt{p}$ donde p es el número inicial de parámetros -, los árboles obtenidos son posteriormente ponderados a partir de su poder predictivo. La independencia entre los diferentes arboles permite al conjunto reducir la varianza.

- **Adaboost:** Corresponde a un método de boosting en donde cada árbol de decisión se entrena basado en los resultados obtenidos anteriormente; por ello depende de un parámetro de aprendizaje el cual es determinado mediante la función de perdida.

- **Xtreme Gradient Boosting:** Es una generalización de Gradient Boosting en la cual, además de generar una función de pérdida distinta en cada árbol utilizado, controla el sobre ajuste al introducir un termino de penalización adicional que funciona como termino de poda. Para más detalles se puede consultar [6] o la pagina de referencia XGboost.

El uso de árboles de decisión es común en la implementación de estos modelos, aunque el método base no se limita a ellos o a la aplicación de un único método, una aplicación de combinación de métodos se puede observar en [5]

2.3. Otras técnicas de clasificación

Otros de los métodos de clasificación que se considerarán en esta tesis son las Maquinas de soporte de Vectores (SVM), las redes Neuronales (Nnet) y los árboles de decisión (DT).

Para SVM se busca determinar un hiperplano separador que maximice la distancia entre las observaciones a clasificar, las Nnet corresponde a una conceptualización de aprendizaje por repetición, mientras que DT corresponden a la base estándar de los métodos de emsemble, dada su fácil interpretación así como implementación.

2.3.1. Árboles de decisión (DT)

Es un algoritmo de aprendizaje supervisado basado en el modelo de partición recursiva. Desarrollado inicialmente por Breiman L et al.(1984) los árboles de decisión pueden utilizarse para modelar tanto un proceso de regresión como de clasificación, logrando trabajar con datos numéricos o categóricos.

En este contexto, se llama nodo de decisión a la variable que permite separar a los datos. En general se utiliza una función de "impureza" f la cual se calcula para cada nodo A y para cada clase i de A. En el caso de la función *rpart* presente en el *R* [47] se tienen como criterio de separación diferentes opciones:

- Indice de impureza de Gini. $f(p_{iA}) = p_i(1 - p_i)$ donde p_i corresponde a la probabilidad de que la clase i sea clasificada de forma incorrecta.
- Ganancia de la información. $f(p_{iA}) = -p_i \log_2(p_i)$ la cual mide la entropia que se puede interpretar como un indice de heterogeneidad entre los datos.
- Error de clasificación. $f(p_{iA}) = 1 - p_i$

Para cada uno de los anteriores se busca maximizar la función de reducción de impureza

$$I(A) = \sum f(p_{iA})$$

Las cuales en el caso binario son equivalentes. El criterio de parada en la creación del árbol dependerá de la profundidad, cantidad de split en cada nodo, o de la separabilidad completa de los datos.

2.3.2. Maquinas de soporte de vectores (SVM)

Dado un conjunto de entrenamiento de m observaciones con p predictores y con variable dependiente y dicotómica, se desea determinar el hiperplano

$$f(x) = X\beta + \beta_0 = 0 \tag{2.7}$$

tal que la distancia entre los conjunto que separa sea máxima. La función de clasificación $G(X)$ se establece redefiniendo la variable respuesta $y \in \{-1, 1\}$ - de ser el caso- como

$$G(x) = sign\,(y \cdot f(x)) \tag{2.8}$$

en donde la función a maximizar será el margen de diferencia entre el valor real y el plano separador. Para esto se toman aquellos valores de β tal que $||\beta|| = 1$.

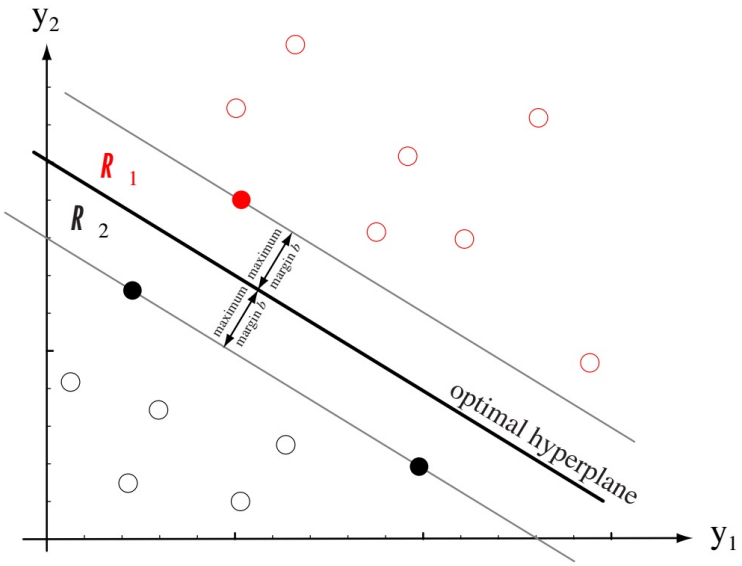


Figura 2.1: El hiperplano separador óptimo corresponde a aquel que maximice las distancias entre los datos que separa. Tomado de [12, cap 5]

2.3.3. Redes neuronales (Nnet)

La idea central de las redes neuronales es establecer combinaciones lineales de las entradas que funcionen como características que permitan describir un modelo no lineal. El número de entradas y salidas es determinado por la cantidad de parámetros x_i y variables respuesta y_i del problema; sin embargo el número de capas ocultas o intermedias y la cantidad de estas son parámetros libres que pueden afectar la calidad de predicción . El paso entre capas es controlado mediante una función de "activación" f_k la cual depende del problema; sin embargo en problemas de clasificación puede ser una función logística o sigmoidal; cada uno de estos pasos es ponderado por un peso w_jk el cual depende de la importancia relativa de la capa JK.

$$z_k = f_k\left(\alpha_k + \sum_{j \to k} w_{jk} f_j \left(\alpha_j + \sum_{i \to j} w_{ij} x_i\right)\right) \tag{2.9}$$

Para el caso de una red de una sola capa oculta con j nodos, i entradas y k salidas, con función de activación f y parámetros de ajuste α. Una red neuronal puede ser modelada como una composición de 2 funciones lineales asociadas a dos funciones de activación (Ecuación 2.9). La Figura 2.2 muestra la arquitectura de este tipo de red

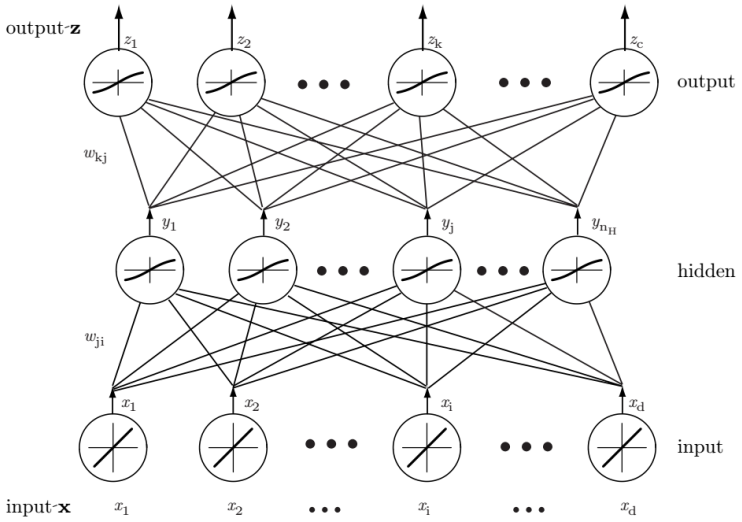


Figura 2.2: Arquitectura de una red neuronal. Tomado de [12, pag 342]

En donde los pesos $w.$ son actualizados acorde a una función de aprendizaje definida por el error de entrenamiento.

2.4. Evaluación de los modelos

En el desarrollo de modelos que ajusten a los datos en cuestión, puede ocurrir que varios cumplan dicha característica, al corresponder a datos multivariados de un modelo de clasificación, se considera como criterio de comparación el uso de la devianza, así como el criterio de información de akaike y el área bajo la curva ROC.

La decisión de un modelo sobre los otros dependerá de:

- La capacidad descriptiva e interpretativa de los datos, por parte del modelo.
- El balance entre la capacidad descriptiva vs número de variables, siguiendo el principio de parsimonia.
- El criterio profesional de los especialistas consultados.

2.4.1. Devianza

Se define la devianza de un modelo como

$$D(y;\mu) = -2[l(\hat{\mu};y) - l(y;y)] \tag{2.10}$$

donde $l(y;y)$ corresponde al logaritmo de la verosimilitud (log-verosimilitud) del modelo saturado vs el modelo propuesto $l(\hat{\mu};y)$.

Como se puede observar $D \to 0$ conforme $l(\hat{\mu};y) \to l(y;y)$ por tanto a menor devianza más se estarán aproximado los resultados de un modelo a los de un modelo saturado.

Por otro lado, al comparar dos modelos entre sí -supongamos con estimadores μ_0 y μ_1 respectivamente- para un mismo conjunto de datos se tiene

$$\begin{aligned} d &= -2[L(\hat{\mu_0};y) - L(\mu_1;y)] \\ &= -2[L(\hat{\mu_0};y) - L(y;y)] + 2[L(\hat{\mu_1};y) - L(y;y)] \\ &= D(y;\mu_0) - D(y;\mu_1) \end{aligned} \tag{2.11}$$

lo que corresponde a una diferencia de las devianzas, las cuales bajo condiciones de regularidad cumple

$$d \sim \chi^2$$

con igual grados de libertad que la diferencia entre el número de parámetros de los modelos, y con hipótesis nula de igualdad. Es decir d establece una prueba estadística para validar igualdad entre modelos.

2.4.2. Criterio de Akaike

El *criterio de información de Akaike* (AIC), corresponde a un indicador sobre la calidad de ajuste de los valores predichos a los reales mediante la relación

$$AIC = -2l(\hat{\mu}_0;y) + 2k \tag{2.12}$$

donde $l(\hat{\mu}_0;y)$ corresponde al máximo valor de la función de log - verosimilitud del modelo estimado y k corresponde al número de parámetros del modelo. Como se observa a mayor número de parámetros se obtiene un mayor AIC por tanto ante dos modelos equivalentes se escoge aquel que tenga un menor AIC.

Es importante destacar que de acuerdo con [1, pág 142] para datos categóricos, la selección de un modelo a partir de la devianza o a de AIC es equivalente.

2.4.3. Curvas ROC

Una recepción operativa de características (ROC) *"...es una técnica de visualización, organización y selección de un clasificador basado en su desempeño..."*[13]. Son utilizadas para evaluar la calidad del diagnostico realizado.

El diseño de una curva ROC depende de una matriz de confusión, la cual se elabora a partir de una tabla de contingencia 2×2 de los valores reales vs los predichos por el modelo propuesto.

	Predicción	
	Positivos	**Negativos**
Positivos	Verdaderos Positivos (VP)	Falsos Positivos (FP)
Negativos	Falsos Negativos (FN)	Verdaderos Negativos (VN)
$\sum$	P	N

Cuadro 2.1: Matriz de confusión.

Para la matriz mostrada en el Cuadro 2.1 se pueden determinar algunos indicadores comunes, los cuales se muestran en el Cuadro 2.2

Para la generación de la curva son particularmente utilizadas la **Sensitividad** y la **Especificidad** como valores de referencia en los ejes X e Y. A mayor cercanía del área bajo la curva (AUC) a uno se dice que es mejor el estimador.

Otros indicador que surge es la **kappa de Cohen**, conocido frecuentemente como estimador **Kappa**, el cual da un indice de *:"que tan probable es que los resultados sean generados por acción del azar P_{azar}, respecto a que sean producto del modelo $P_{acuerdo}$"*.[1]

El estimador

$$\kappa = \frac{p_{azar} - p_{acuerdo}}{1 - p_{acuerdo}}$$

indica que valores cercanos a cero corresponden a resultados completamente obtenidos por azar, mientras que valores cercanos a uno a resultados obtenidos mediante la predicción de forma adecuada.

[1] Usualmente se utiliza para comparar criterios; motivo por el cual en lugar de modelo se suele utilizar "acuerdo"

Precisión: Indica que porcentaje de los clasificados positivos son verdaderamente positivos $$precision = \frac{VP}{VP + FP}$$	**Exactitud** : Indica el porcentaje de datos clasificados correctamente $$Exactitud = \frac{VP + VN}{P + N}$$
Tasa de error: Porcentaje de datos clasificados incorrectamente $$\text{Tasa de Error} = \frac{FP + FN}{P + N}$$	**Sensitividad** : También conocido como tasa de positivos (True Positive Rate), corresponde al porcentaje de datos positivos clasificados correctamente $$Sensitividad = \frac{VP}{VP + FN} = \frac{VP}{P}$$ *Equivalente a la potencia o poder estadístico
Especificidad: porcentaje de datos negativos clasificados correctamente, también conocido como tasa de negativos correctos (True Negative Rate) $$Especificidad = \frac{VN}{FP + VN} = \frac{VN}{N}$$	**Medida - F**: Corresponde a la media armónica de la precisión y la sensitividad. Es una medida de precisión que indica una precisión y sensitividad perfecta en 1 y el peor escenario en 0 $$F - measure = 2 \cdot \frac{precision \times sensitividad}{precision + sensitividad}$$

Cuadro 2.2: Indices que surgen de la matriz de confusión.

2.5. Asociación entre variables

Ya sea en un modelo de clasificación o de regresión, la **dependencia** entre variables predictivas puede provocar inflación de la varianza, disminución en el poder predictivo, falso ajuste, aumento del coste computacional y económico de la implementación del modelo, entre otros. por tanto, la independencia entre variables es un rasgo deseable en la determinación de un modelo adecuado.

Algunas formas de medir dependencia son:

2.5.1. Correlación Tetracórica

Para un conjunto de variables categóricas dicotómicas Freiberg Hoffmann et al. [17] indica que se recomienda el uso de la correlación tetracórica para describir su relación r, por encima del uso generalizado del coeficiente de Pearson. Esto, al conservar de mejor manera, las relaciones no lineales entre las variables.

El coeficiente de correlación **r** es determinado mediante cálculo recursivo de la función inversa 2.13,

$$L(h,k,r) = \frac{1}{2\pi\sqrt{1-r^2}} \int_k^\infty \int_h^\infty \exp\big(-\frac{x^2+y^2-2xy}{2(1-r^2)}\big)dxdy \tag{2.13}$$

Los detalles de la solución se pueden encontrar en [10]

2.5.2. Escalamiento Multidimensional (MDS)

Corresponde a una técnica en la que a partir de una matriz de distancias se realizan transformaciones para intentar preservar las asociaciones en una dimensión menor a la original.

Suponga una matriz de datos **X** expresada mediante vectores fila **p** dimensionales

$$\begin{aligned} X_1 &= \{x_{11}, \cdots, x_{1p}\} \\ &\vdots \\ X_n &= \{x_{n1}, \cdots, x_{np}\} \end{aligned}$$

De la matriz de datos se define la matriz de distancias, la cual en el caso de utilizarse la distancia euclidea se expresa como $d_{ij} = |X_j - X_i||$. La evaluación del modelo es a partir del estrés **S** de ajuste con las nuevas distancias d^*_{ij}

$$S = \sum_i^n \sum_j^n \big(d_{ij}^2 - (d_{ij}^*)^2\big)$$

Para esto se debe calcular la nueva matriz de distancias, a partir del criterio de ajuste deseado, al respecto $d^*_{ij} = V \cdot D \cdot V^t$ la cual corresponde a la factorización de la matriz $\mathbf{XX^t}$ en su forma ortonormal, donde **D** es una matriz diagonal de valores propios que representan la retención de la varianza y **V** corresponde a los respectivos vectores propios. Más información al respecto, se puede encontrar en [25].

2.5.3. Análisis de componentes principales (PCA)

Dado un conjunto de variables aleatorias $X = \{x_1, \cdots, x_p\}$ que representen los predictores de una expresión $Y = X\beta + \epsilon$, sin pérdida de generalidad suponga que cada predictor esta estandarizado -media $\mu = 0$ y varianza $\sigma^2 = 1$ – .-

PCA corresponde a una técnica que consiste en la creación de nuevas variables $U = \{u_1, \cdots, u_k\}$ a partir de la combinación lineal de las variables de X

$$\begin{aligned} u_1 &= a_{11}x_1 + \cdots + a_{1p}x_p \\ &\vdots \\ u_k &= a_{k1}x_1 + \cdots + a_{kp}x_p \end{aligned}$$

La cantidad de variables **k** a utilizar dependerá de la varianza que se desee mantener de los datos, esta ultima determinada a partir de los eigenvalores α_i con $i = 1, \cdots, p$ de la matriz de covarianza

$$Cov(X) = X^tX = \alpha_1V_1 + \cdots + \alpha_nV_n \approx \alpha_1V_1 + \cdots + \alpha_kV_k = U^tU$$

Donde V_i corresponden las matrices asociadas a cada eigenvector que determinan la descomposición espectral de X^tX. La varianza considera es dada por el número de valores propios considerados.

2.5.4. Análisis de correspondencia múltiple(MCA)

Considere una matriz binaria X de n observaciones y p variables representada $X = \{X_1, \cdots, X_n\}^t$. Defina la matriz de Burt como

$$Z = X^tX = \begin{pmatrix} X_1^tX_1 & \cdots & X_1^tX_n \\ \vdots & \vdots & \vdots \\ X_n^tX_1 & \cdots & X_n^tX_n \end{pmatrix}$$

La tabulación de las variables i, j están representadas por $X_i^tX_j$ en donde $X_i^tX_i$ es una matriz diagonal. La matriz Z es simétrica, la cual cumple

$$\lambda_k^Z = \lambda_k^{Z^t}$$

Lo cual implica que la proyección sobre las frecuencias de las observaciones tiene los mismos eigenvalores que la proyección sobre las frecuencias de las variables. En estos términos, se observa que λ no representa una varianza como tal, al corresponder a resultados originados a partir de una tabla de frecuencia y no a una matriz de covarianza, por ende se utiliza el concepto de *"inercia"* el cual refiere a la capacidad de observación del cambio en los datos. Para mayor detalle consultar [8, p173].

2.5.5. Test χ^2 de Pearson

Utilizado para contrastar independencia entre proporciones de variables categóricas. [1]Se basa en el estadístico χ^2 para determinar la asociación entre las variables El

[1]Cuando el tamaño de muestra es pequeño se recomienda el test exacto de Fisher.

estadístico esta determinado por

$$\chi^2 = \sum_j \sum_i \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{2.14}$$

donde μ_{ij} corresponde al valor esperado y se aproxima por

$$\mu_{ij} = \frac{n_{i+} \times n_{j+}}{n}$$

con n_i el total por fila, n_j el total por columna y n el total de observaciones. Se considera como grados de libertad

$$df = (numfilas - 1) \times (numcolumn - 1)$$

bajo la hipótesis nula de independencia. Para comparar el grado de asociación entre dos o más tablas de contingencia se puede utilizar la V de Cramer, el cual es un valor entre 0 y 1 que indica que valores cercanos a 1 tienen una mayor dependencia. El índice se calcula

$$V = \sqrt{\frac{\chi^2}{n \cdot m}} \tag{2.15}$$

con $m = min(numfil - 1, numcol - 1)$

Esta relación será utilizada posteriormente en el Capítulo 4 para establecer asociaciones entre las variables e indicar la fuerza de asociación acorde a la magnitud del estadístico.

2.5.6. Momios o Razón de la ventaja (OR)

Conocido también como *odds ratio*, los OR representan el porcentaje de beneficio de presentar la característica asociada a la probabilidad p, respecto a no presentarla [41].

En el caso de un modelo logístico como el indicado en 2.1, la comparativa es inmediata al seleccionar las variables de interés X_i del modelo se obtiene una probabilidad p_i, la cual cumple $y_i = log\left(\frac{p_i}{1-p_i}\right)$, luego

$$OR(X_i) = e^{y_i} = \frac{p_i}{1 - p_i} = e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_p x_p} \tag{2.16}$$

De forma similar, en el caso de tablas de contingencia 2×2 para variables dicotomicas -como la presentada en el Cuadro 2.3 - se tiene que los OR se pueden calcular mediante la relación

$$OR(var_i) = \frac{a \times d}{c \times b} \tag{2.17}$$

	Caso	Control	total
$var_i = 1$ (p)	a	b	$a+b$
$var_i = 0$ (1-p)	c	d	$c+d$
total	a+c	b+d	n

Cuadro 2.3: OR a partir de una tabla de contingencia 2×2

Respecto al uso de tablas de contingencia Agresti [pag 71, 1] indica que en el caso de que alguna entrada sea nula, se puede realizar la corrección:

$$OR(P) = \theta = \frac{(a+0{,}5) \times (d+0{,}5)}{(c+0{,}5) \times (d+0{,}5)} \tag{2.18}$$

Lo cual evita tener valores a infinito o nulos.

De acuerdo con [37] el uso frecuente que se da a los *odd ratio* en las áreas de la salud estableciéndolo como indicador de las tasas de incidencia permite dar las siguientes interpretaciones a los valores de OR:

- $OR < 1$ La variable tiene un efecto a favor del caso considerado
- $OR = 1$ La exposición a la variable no presenta asociación observable al caso en estudio.
- $OR > 1$ La exposición a la variable aumenta las posibilidades contra el caso considerado.

Luego, considerando la relevancia que tienen la cercanía de los OR respecto al 1, se puede establecer los intervalos de confianza generados a partir de las tablas de contingencia (IC de Wald) mediante la relación

$$IC = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.19}$$

donde $z_{\alpha/2}$ corresponde al nivel de significancia de una distribución normal [ver 1, pag 185]. En el caso de que $\hat{p} = 1$, entonces el mismo autor señala se puede establecer la corrección $\log OR(p)$ y generar los intervalos de confianza a partir aplicar la función exponencial a la ecuación (2.20)

$$\log \hat{\theta} \pm z_{\alpha/2} \hat{\sigma}\big(\log \hat{\theta}\big) \tag{2.20}$$

donde $\hat{\sigma}\big(\log \hat{\theta}\big) = \big(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\big)^{1/2}$

En el caso del modelo logístico, los IC se determinan a partir de los intervalos de confianza de los coeficientes utilizados en el calculo del OR.

2.6. Imputación

Un método de imputación, corresponde a un método predictivo el cual basado en los datos presentes "predice" los datos poco informativos en las diferentes variables.

Para realizar imputación de forma adecuada por lo general se pide que los datos faltantes sean completamente aleatorios (MCAR), es decir que la probabilidad de que un dato de una observación sea faltante, sea equivalente para todos los valores faltantes que corresponden a una matriz de datos $X = \{X_{ob}, X_{per}\}$, Donde X_{ob}, X_{per} corresponden a la parte observada y no observable respectivamente; además se debe cumplir que X_{per} se den de forma independiente respecto a los X_{ob}

$$P(Z|X_{ob}, X_{per}, \tau) = P(Z, \tau) \tag{2.21}$$

En la ecuación (2.21) τ corresponde a los paramentos desconocidos que modelan a la distribución Z de los datos. Sin embargo este hecho no siempre se presenta y es más natural que la naturaleza de los datos faltantes dependa de las variables observadas -en este caso se dice que solo son aleatorios o MAR -. como se indica en la ecuación (2.22)

$$P(Z|X_{ob}, X_{per}, \tau) = P(Z|X_{ob}, \tau) \tag{2.22}$$

Cuando no se cumple la condición de MAR o MCAR, se corre el riesgo de que al imputar se introduzcan sesgo que distorsionen la naturaleza de los datos, el método de imputación múltiple corresponde al proceso de imputar de forma reiterada e independiente el mismo dato y combinar los resultados de forma tal que se obtenga una estimación insesgada. Detalles del proceso se pueden consultar en[16, 28, 48].

Capítulo 3

Características de las bases de datos

En el presente estudio se utilizaron dos bases de datos correspondientes a diferentes fases del proyecto CHIBCHA; la *Base 1* refiere a datos clínicos de pacientes mexicanos a los cuales se les tomaron muestras de sangre y se les aplico una entrevista (presente en el anexo), mientras que la *Base 2* corresponde a un refinamiento de la primera, en la cual se selecciona a individuos para realizar un genotipado.

La codificación de los datos de cada base es categórica, siendo la variable *edad* la única de tipo discreta; para la cual en el presente estudio se realiza una re-codificación para que sea trabajada como categórica.

Para los datos faltantes de cada *Base* se realizó imputación múltiple; en el caso de la *Base 1* mediante el método de arboles de decisión *cart* de la librería *mice* [48], mientras que para la *Base 2* mediante el algoritmo de EM (Expectation–Maximization) presente en el paquete *missMDA* mediante la función *imputeMCA* [28]. Los métodos se seleccionaron por conveniencia, al mostrar mejor calidad predictiva en cada base de datos.

La *Base 1* tiene 13 variables con hasta 7 niveles y 3525 observaciones correspondientes a la codificación de las entrevistas aplicadas por el equipo de investigación de México; mientras que la *Base 2* engloba a 77 variables de hasta 3 niveles correspondientes a SNP identificados como significativos por el equipo de investigación de Uruguay - España.

Ambas bases se unieron utilizando la etiqueta de individuo, el genero y el fenotipo, variables que fueron comunes en ambas fases, generando una *Base Final* con 1707 observaciones -879 casos y 828 controles- y 89 variables (204 al pasar a *dummy* las diferentes categorías). A continuación se describen las características de cada una de las bases y su proceso de estructuración a la *Base Final*.

3.1. Variables procedentes de la Fase 1 (Base 1)

Producto de la coordinación del proyecto CHIBCHA para México, se recopiló información de 3525 personas relativa a genero, edad, nivel educativo, existencia de familiares con CCR, tenencia de pólipos, obesidad, sedentarismo, presencia de diabetes, fumado, alcoholismo y consumo de carne. La Tabla 3.1 ahonda en lo anterior.

Variable de la base	Escala	Descripción
Age	Discreta	Edad en años cumplidos
Fenotype (Fenot)	1,0	0 - presenta CCR (caso) o 1- no presenta CCR (control)
Gender(Gene)	1,0	0- Representa mujer, 1- Representa hombre.
Familywith ColonCancer (FWCC)	1,0	1-Si tiene familiares con CCR otros familiares con CCR diagnosticado, 0- otro caso.
Polypus (Pol)	1,0	1- Si se le han detectado pólipos con anterioridad , 0 - Otro caso
Obesity (Obe)	1,0	A partir de la talla y peso se realiza un cálculo del Indice de Masa Corporal (IMC) y se anota el estatus 0- sin sobre peso, 1- Con sobre peso
Sedentariness (Sed)	1,0	Realizan actividad física 0- no realiza 1- realiza
Diabetes (Diab)	1,0	1- padece de diabetes, 0- no padece
Family with Extra Colon Cancer	1,0	1- Si tiene familiares que tienen otros tipos de cáncer diagnosticados y se anota el número de tipos de cáncer diagnosticados en la familia. Se selecciona y 0 - si no los tiene
Smoking (Smok)	0,1,2,3	Hábitos de fumado: 0-No fuma, 1-Menos de 5 cigarrillos por día, 2 de 5-20 cigarrillos por día, 3 más de 20 cigarrillos por dia
Alcohol Consumption (Alc)	0,1,2,3	Consumo de alcohol: 0- No consume, 1 menos de 75mL a la semana, 2 entre 75-150ML a la semana, 3 más de 150mL a la semana
Educational Level (Ed)	0,1,2,3,4,5,6	0- sin escolaridad, 1 - primaria completa, 2-secundaria completa, 3-preparatoria, 4-Estudios Técnicos, 5-Estudios Universitarios, 6- Estudios de posgrado.
Red meat intake (Carne)	0,1,2,3	Frecuencia de consumo de carnes rojas, se establecen 4 niveles: 0 - no consume, 1- menos de 1500g por semana, 2- entre 1500-4500 gramos por semana, 3 - mas de 4500 gramos por semana.

Cuadro 3.1: Descripción de las variables *Base 1.*

En la *Base 1* se tiene información de 2091 (59 %) controles y fue recopilada mediante muestras del banco de sangre tomadas por conveniencia a pacientes sanos. Mientras que 1434 (41 %) casos se tomaron mediante entrevistas realizadas a los pacientes ya diagnosticados.

	Centro Médico	% de participación
CDMX	Centros de Salud de Ciudad de México.	74.6 %
UANL	Centros de Salud de la UANL.	9.2 %
IMSS-MTY	Centros de Salud de Monterrey.	8.9 %
BSH	Laboratorio Clínico BSH (Monterrey).	5.7 %
Torreon	Distintos Centros de Salud Torreón.	1.6 %

Cuadro 3.2: Distribución de la muestra por centro médico.

La recolección de la información se da entre 2010 - 2013 y corresponden a distintas locaciones mexicanas, principalmente de Ciudad de México y Nuevo León - Monterrey como se aprecia en el Cuadro 3.2.

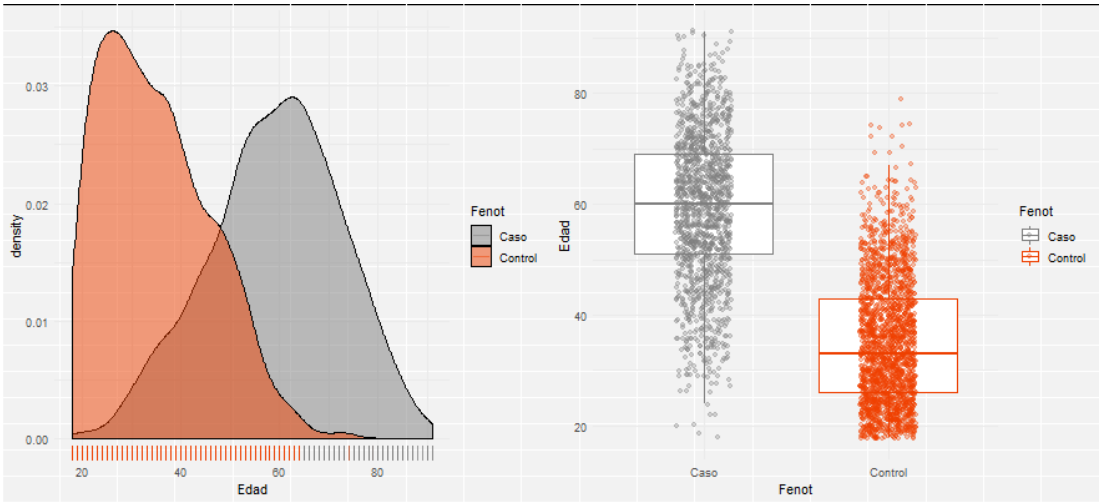


Figura 3.1: Distribución de los cohortes.

Note que si bien todos los individuos pertenecen a México las características regionales de cada estado podría conducir a un sesgo, producto del desbalance en la toma de los datos. Otro aspecto a considerar es el desbalance entre los casos y controles; para la variable *"Edad"*, la Figura 3.1 permite observar mediante el gráfico de densidad un sesgo a la izquierda en los controles, el mismo que se refleja en el gráfico de caja lo cual repercute en el efecto separador por parte de la edad.

3.1.1. Imputación para la Base 1

Se presentaron 282 datos faltantes los cuales fueron imputados mediante el método de imputación múltiple presente en el paquete *mice* [48] y que utiliza un algoritmo de árboles de decisión.

La Figura 3.2 permite observar que la variable *"carne"* fue la que más datos faltantes presentó con aproximadamente 2.5 % (96 datos) del total. El gráfico de barras de la izquierda muestra la frecuencia relativa de las variables faltantes dentro de todo el conjunto; mientras que el cuadro de la derecha muestra los datos faltantes para las diferentes combinaciones de variables, los valores relativos se muestran al lado derecho. Aproximadamente el 94 % de los datos no presentaba faltantes de información.

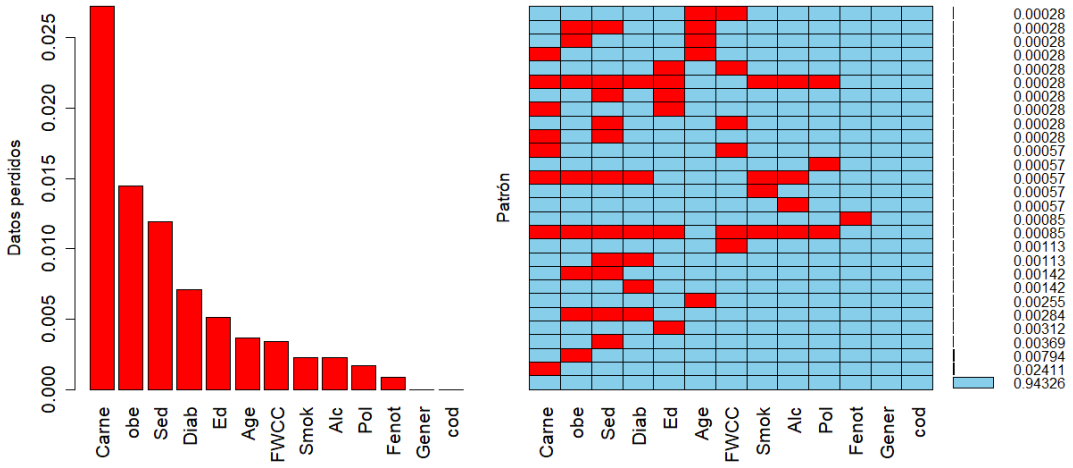


Figura 3.2: Patrón de datos perdidos en la Base 1.

La calidad de la imputación se determinó a partir de la amputación de un 5 % de los datos, los cuales posteriormente fueron imputados mediante el mismo método. Se observo un error relativo de 0.17 % en las predicciones.

3.1.2. Grupos de edad

Para la variable *"edad"* se realiza una categorización mediante el algoritmo K-means, el cual busca determinar el número de grupos **G** que minimice la suma de las distancias de cada dato respecto a su media. Es decir se desea calcular el mínimo de

$$argmin \sum_{i=1}^{k} \sum_{x_j \in G_i} \sqrt{||x_j - \bar{x_i}||^2} \tag{3.1}$$

Se utiliza la opción predeterminada del programa *R* presente en la paqueteria *stats* mediante el comando *kmeans*, la cual usa el algoritmo de Hartigan [46] para realizar las agrupaciones. Al considerar como limite que cada grupo tenga al menos un 1 % de los datos, se determina que 5 agrupaciones son adecuadas explicando un 95 % de la varianza la Figura 3.3 muestra los el error cuadrático de realizar esta agrupación al lado izquierdo, mientras que a la derecha la explicación de la varianza.

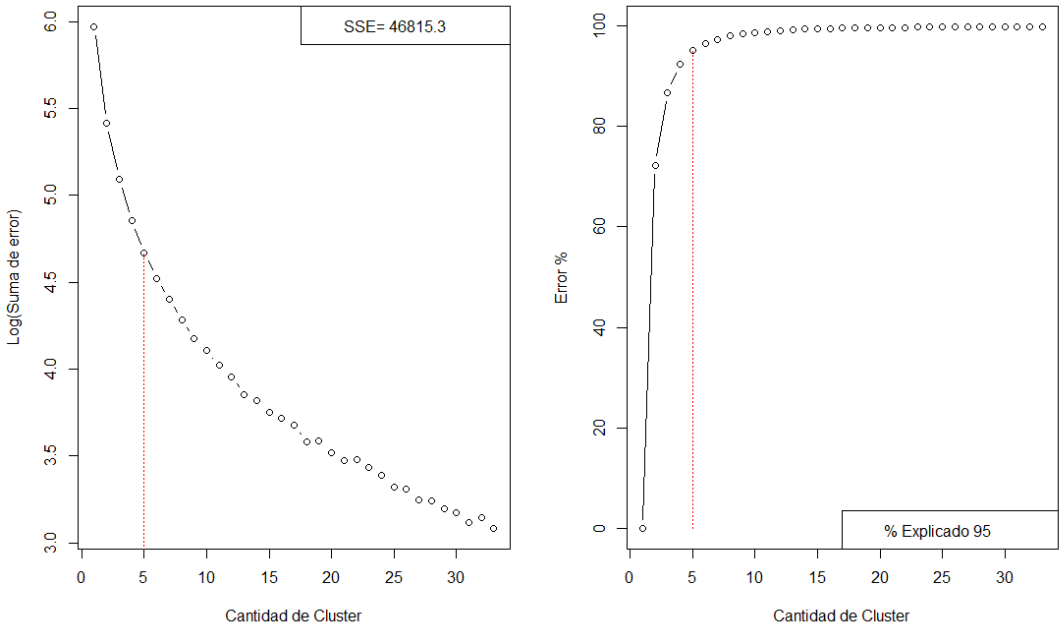


Figura 3.3: Creación de los grupos de edad.

La agrupación de las edades permite crear la nueva variable *"Nage"* la cual incluye 5 niveles cuyos límites se muestran en el Cuadro 3.3.

Se prefirió esta categorización sobre la clasificación por quintiles con la intención de mantener la estructura global de los datos y no la homogeneidad de los grupos. Esto se considero más adecuado para describir el comportamiento original, aunque en ambos casos los resultados son similares. Al resumir la información de la muestra a partir de la categorización de la edad, se obtiene la Figura 3.4 en donde se observa la presencia intrínseca de dos grupos de variables, la linea horizontal de umbral seleccionada al 41 % -correspondiente a la proporción de casos- muestra desproporciones en algunas de las categorías, lo que representa un problema al desarrollar algunos de los modelos de clasificación de la sección 5, esto dada la poca variabilidad que aportan.

	Nage1 (18-29)	Nage2 (30-41)	Nage3 (42-54)	Nage4 (55-68)	Nage5 (69 o más)
Casos	0.65	3.69	10.21	15.74	10.41
Controles	22.35	20.62	13.28	2.78	0.26
Total aproximado	23 %	19 %	20 %	22 %	17 %
	Qage1 (18-28)	**Qage2 (29-38)**	**Qage3 (39-49)**	**Qage4 (50-61)**	**Qage5 (62 o más)**
Caso	0.43	2.75	6.07	12.88	18.55
Control	20.48	18.07	13.56	6.35	0.85
Total aproximado	21 %	21 %	20 %	19 %	19 %

Cuadro 3.3: Agrupación de las edades para los 3525 datos, se muestra la comparativa entre las quintiles (Qage) y los grupos de edad (Nage).

Otro rasgo de la Figura 3.4 corresponde a la tendencia de decrecimiento de los casos respecto a los controles en los grupos de edad *"Nage"*, hecho que posteriormente coincide con su fuerte efecto clasificador.

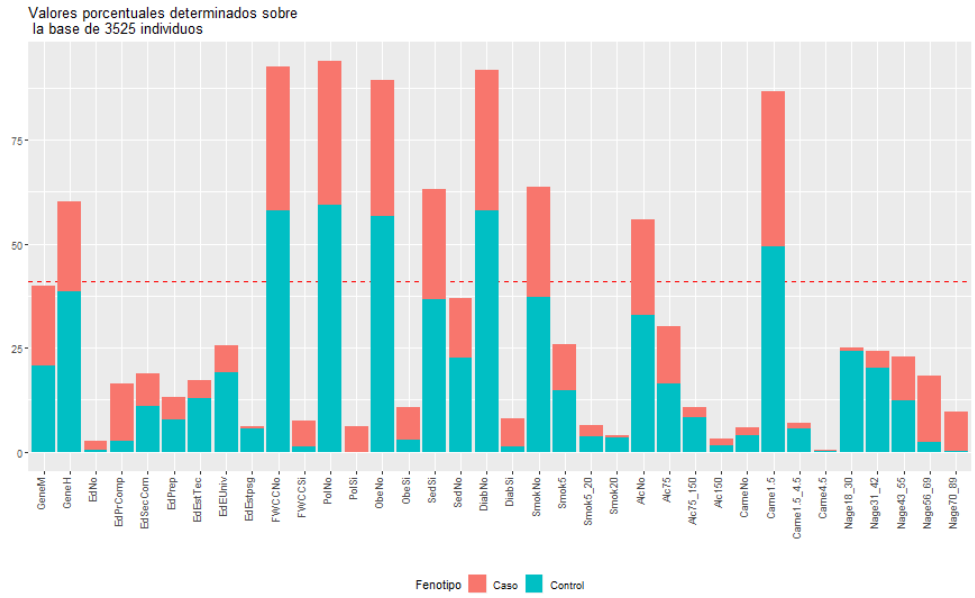


Figura 3.4: Frecuencias relativas de las variables correspondientes a la *Base 1* por fenotipo.

3.1.3. Asociaciones entre las variables

Dada la naturaleza categórica de los datos se propone el uso del coeficiente de correlación tetracórica, el cual de acuerdo a [17] describe de manera más adecuada las correlaciones al considerar las categorías de cada clase, respecto al coeficiente de Pearson el cual las supone independientes. El cálculo se realiza mediante la función *polycor* presente en el paquete *psych* [39] cuyas justificaciones teóricas se encuentran en [10, 30]. Los resultados se muestran en la Figura 3.5.

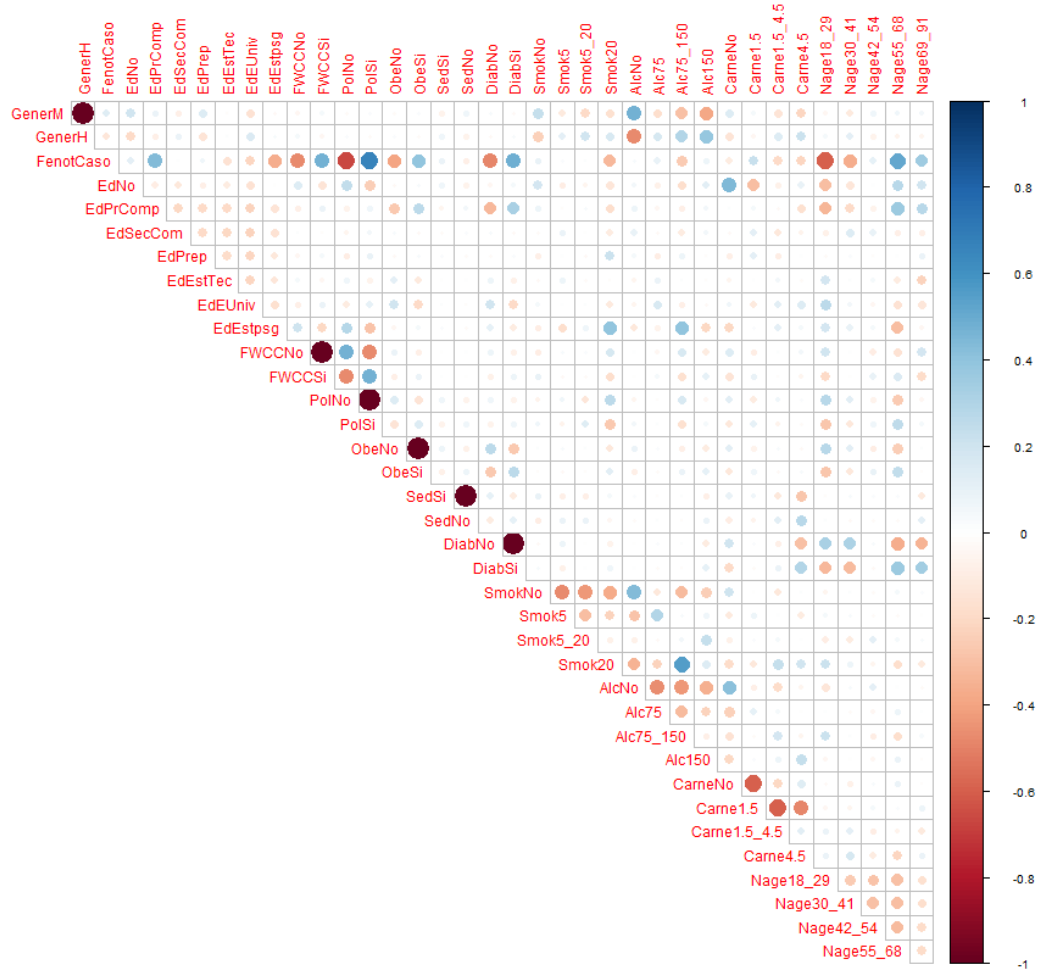


Figura 3.5: Correlación tetracórica entre las variables de la Base 1.

Del mismo modo, las relaciones entre variables como *EdNo* $\sim$ *Carne*, *EdEstPosg*

$\sim$ *Smok20*, *Alc75-150*, *SmokNo*$\sim$*AlcNo*, *Diab*$\sim$*Nage+55* son de particular interés al corresponder a posibles agrupaciones de para un mismo perfil las cuales al crear dependencias generan problemas de multicolinealidad.

Una forma de analizar con más detalle las dependencias descritas anteriormente, corresponde a realizar una proyección en componentes principales (PCA).

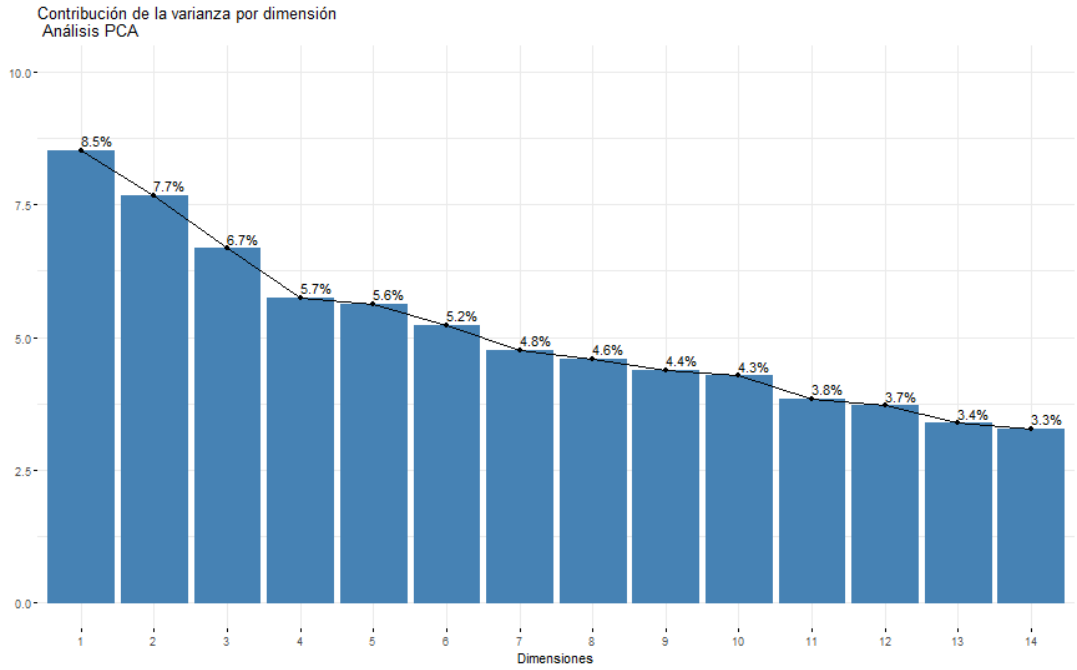


Figura 3.6: La alta variabilidad de los datos se muestra en la distribución de la varianza.

22 dimensiones abarcan alrededor de un 94 % de la información de las cuales las tres primeras corresponden a poco más del 22 % de la varianza - ver Figura 3.6- lo que hace que el método sea poco práctica para establecer relaciones interpretables. El escalamiento multidimensional (ver [25, Cap 14]) por su parte, permite realizar una primera aproximación a relaciones entre variables de la *Base 1* al mostrar sus proximidades de acuerdo a la distancia de Gower. El escalamiento es aplicado mediante la función *cmdscale* del paquete base *MASS* [49].

Se tiene que con 22 dimensiones, el estres baja a 0.2; a partir de estas se aplica k-means para establecer los 5 grupos mostrados; los cuales representan el 96 % de la varianza de los datos. En la Figura 3.7 los datos están separados por *Fenotipo* y utilizan la misma codificación de color para cada grupo, se observa como se da una variación marcada en las variables del grupo 2 lo que corresponde con considerarlas variables diferenciadoras entre los casos y los controles.

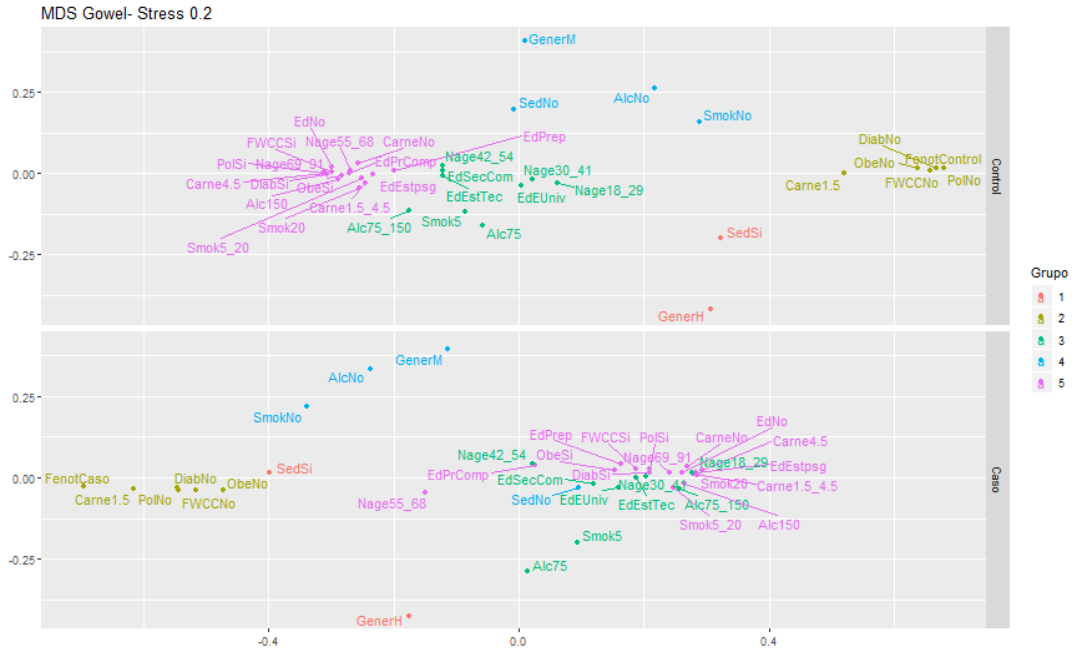


Figura 3.7: Al considerar 5 grupos se obtiene una explicación del 94 % de la información.

A pesar del alto estrés del método, se observa que la proyección corresponden a un punto de inicio para los análisis posteriores al mostrar variables con comportamiento similar las cuales podrían corresponder con un perfil para los individuos.

3.2. Variables procedentes de la Fase 2 (Base 2)

Posterior a la recopilación de las 3525 observaciones de la Fase 1 (Base 1) correspondió al equipo de investigación de Uruguay - España realizar el genotipado de las muestras, las cuales luego de los controles de calidad al que fueron sometidos generaron 1707 observaciones validas distribuidas en 879 controles y 828 casos. La información se analiza en una primera instancia mediante el programa *plink* el cual requiere de los archivos en formato estándar para muestras de pedigree (arhivo.ped) e información de las variantes (archivo.map). La estructura del archivo .ped es mostrada en el Cuadro 3.4 (Se muestran solo 4 columnas del total).

	fam	Cod	P	M	Gener	Fenotipo	1	2	3	4	5	6	7	8
1	FAM2	100	0	0	1	2	C	C	G	G	T	T	A	A
2	FAM3	101	0	0	1	2	C	C	G	G	C	T	A	A
3	FAM4	1010	0	0	1	2	C	C	G	G	T	T	A	A

Cuadro 3.4: Estructura archivo de secuenciación genómica.

La caracterización de cada columna se presenta en el Cuadro 3.5, en el cual se descartan las columnas 2 a 4 al no aportar información de interés al estudio realizado. Esto debido a que los individuos no tienen parentesco entre sí. Las columnas *"Gener"* y *"Fenotipo"* serán utilizadas para validación de la combinación de las bases.

	Identificador del número de muestra 1-1707
fam	identificador de la familia con la que está asociado de acuerdo a la base de *plink* (no se utilizará)
Cod	código de identificador individual
P	identificador del padre
M	identificador de la madre
Gener	1=hombre; 2=Mujer
Fenotipo	1- control, 2 - caso
1 - (en adelante)	corresponde al SNP analizado, se lee por parejas, cada número impar es un snp de cada allelo, el cual se asocia al archivo .map

Cuadro 3.5: Caracterización de las variables de asociación genética.

Cada SNP es nombrado mediante el archivo .map, el cual tiene la estructura mostrada en el Cuadro 3.6 (solo se presentan 4 snp's)

N snp	Crom	snp id-rs	C.morgans	bp
1	1	rs55885037	0.1531	9245640
2	1	rs115829688	0.3754	19984808
3	1	rs74382455	1.0821	84373659
4	1	rs13375831	1.4463	150917632

Cuadro 3.6: Estructura del archivo de etiquetas de los SNP.

Aquí *"N snp"* corresponde al número de SNP en el archivo *.ped*, *"Crom"* al cromosoma autosómico(1-22), *"snp id-rs"* al nombre de utilizado para referirse al SNP, *"C.morgans"* la unidad Centimorgan la cual es utilizada para medir el grado de proximidad genética entre SNP y la probabilidad de que el marcador sea separado de un segundo marcador, mientras que Par de Bases(bp), similar al Centimorgan, corresponde a la cantidad de bases nitrogenadas consecutivas antes de un cambio de base.

N	Crom	snp.id.rs	C.morgans	bp	N	Crom	snp.id.rs	C.morgans	bp	N	Crom	snp.id.rs	C.morgans	bp
1	1	rs55885037	0.1531	9245640	27	6	rs16868695	0.5506	35214884	53	16	rs7197593	0.7024	55405713
2	1	rs115829688	0.3754	19984808	28	6	rs75284034	1.1488	110807192	54	16	rs2194310	0.7827	60069270
3	1	rs74382455	1.0821	84373659	29	6	rs74896351	1.6188	154812412	55	16	rs11860295	0.847	67316234
4	1	rs13375831	1.4463	150917632	30	6	Affx-28073014	1.7002	160557271	56	16	rs3868142	0.847	67320223
5	1	rs74124668	1.5788	161185311	31	7	rs2598121	0.5902	37936286	57	16	rs9922085	0.8471	67397580
6	1	rs76284034	1.8043	180362817	32	7	rs139495018	1.3805	133176331	58	16	rs8047080	0.8471	67402588
7	1	rs74455361	2.3275	229733500	33	7	rs78775244	1.5038	141494751	59	16	rs8052655	0.8471	67409180
8	2	rs116017843	1.194	106723600	34	8	rs2741142	0.161	6739730	60	16	rs7205526	0.8471	67410583
9	2	rs116537234	1.5564	144530334	35	8	rs72473922	0.462	27347878	61	16	rs79915255	1.2703	87862948
10	2	rs2354487	1.7596	170998085	36	8	rs79202010	0.9716	92406834	62	16	rs7192381	1.2713	87905403
11	2	rs115268253	2.0324	206487964	37	8	rs115920066	1.4315	133498971	63	16	Affx-36084964	1.2715	87911526
12	2	rs114152526	2.5895	241541955	38	8	rs76371563	1.4649	134649104	64	17	Affx-13613185	0.0651	1808525
13	3	rs116847323	0.0957	3192669	39	9	rs11568414	0.8232	86907537	65	17	rs114885856	0.1772	6578741
14	3	rs76500230	0.4352	22460064	40	10	rs75514416	1.1924	100153052	66	17	rs116658596	0.9285	61579215
15	3	rs26934	0.8796	64154998	41	11	rs76616706	1.141	114013828	67	18	rs115246312	0.6172	39532320
16	4	rs116770339	0.092	5421128	42	12	rs11054867	0.2817	12536324	68	18	rs76528184	0.6724	45005097
17	4	rs1797044	0.4796	28531139	43	12	rs115132597	0.996	90043331	69	18	rs73955609	0.7445	49866432
18	4	rs74719289	0.8542	77081765	44	12	rs74651174	1.033	93515535	70	18	rs74496677	1.0268	70829140
19	4	rs11935039	1.0177	95380012	45	12	rs34151234	1.3401	117452529	71	19	rs77960487	0.2645	8454998
20	4	Affx-23494624	1.3453	139091425	46	12	rs114474674	1.4123	121682293	72	19	rs1803767	0.3633	14625688
21	4	rs112645305	1.5318	159593199	47	12	rs118184226	1.5014	127302029	73	19	rs76698129	0.5057	30286770
22	5	rs185865591	0.1946	7308316	48	14	rs76646466	0.8132	81977253	74	21	rs73340724	0.259	27406093
23	5	rs192567137	0.681	52183661	49	14	rs73323468	0.8642	88314360	75	21	Affx-18048474	0.5547	42536667
24	5	rs78163585	0.959	79946844	50	14	rs61989760	0.9377	91496423	76	21	rs143772364	0.6039	43716648
25	5	rs11949094	1.8684	172184290	51	14	rs57798805	1.1523	100874491	77	22	rs17002839	0.1764	23955309
26	6	rs115490452	0.4357	21233293	52	16	rs116393639	0.1881	7375875					

Cuadro 3.7: SNP identificados como significativos por el equipo de investigación de Uruguay - España.

El genotipado realizado por el equipo de investigación de Uruguay - España involu-

craba a aproximadamente un millón de SNP, de los cuales se reconocen como relevantes 77, esto de acuerdo a los resultados de utilizar pruebas False Discovery Rate (FDR) e intervalos de confianza de Bonferroni. Los SNP suministrados se muestran en el Cuadro 3.7.

3.2.1. Imputación para la Base 2

Se presentaron 2476 bases nitrogenadas faltantes (del total de 165 088), las cuales son imputadas mediante el algoritmo EM utilizado de forma predeterminada por el paquete *imputeMCA* [28]. El algoritmo es utilizado por conveniencia al darse la imposibilidad de utilizar *mice* dado el formato y naturaleza de los datos.

El paquete *imputeMCA* realiza un análisis de correspondencia múltiple proyectando los datos a un número de dimensiones pre-establecidas; en nuestro caso se utilizaron 5 dimensiones, las cuales explican un 65 % de la variabilidad. La validación de resultados se utilizó generando un conjunto de 5 % de datos MCAR los cuales posteriormente se imputaron. La validación mostró un error relativo de 0.000759 %.

3.2.2. Asociaciones entre los SNP

A partir del programa *Haploview* se establecen las asociaciones entre los diferentes SNP, utilizando el coeficiente r^2 de correlación entre loci, el cual de acuerdo a la documentación de la pagina oficial del proyecto www.broadinstitute.org corresponde a la proximidad entre dos SNP.

Es conocido que el genoma se hereda en bloques [20], en este sentido la proximidad entre SNP aumenta la probabilidad de que se hereden en conjunto. Los estudios de desequilibrio de Ligamiento (LD) -como se menciona en el capítulo 1- estudian estas probabilidades las cuales son basadas en frecuencias biológicas; y no necesariamente en una relación de linealidad como ocurre en el caso del coeficiente de Pearson.

Por lo anterior, el desequilibrio de ligamiento tiene un mayor sentido entre grupos de SNP de un mismo cromosoma dado que físicamente no estarán próximos en cromosomas distintos. Para el conjunto de SNP estudiados se observa una baja correlación -en el sentido biológico- entre SNP de un mismo cromosoma; a excepción de un bloque marcado con los SNP55 -SNP60 correspondientes al cromosoma 16.

3.3. Base Final

Luego de preparadas las *Base 1* y *Base 2* se realiza una combinación del total de datos. Se combinan los alelos de cada SNP para crear variables con hasta 3 niveles

(homocigoto dominante, heterocigoto, homocigoto recesivo) generando una base que contiene 89 variables y 1707 observaciones.

Obs	Gener	Fenot	Ed	FWCC	Pol	Obe	Sed	Diab	Smok	Alc	Carne	Nage	rs55885037.1
1	1	0	2	1	1	1	1	0	0	2	1	2	CC
2	0	1	3	0	0	0	0	0	0	2	1	2	CC
3	0	1	3	0	0	1	1	0	0	2	1	3	CC
4	1	1	3	0	0	0	0	0	3	2	1	1	CG
5	1	1	2	0	0	0	0	0	3	2	1	2	CC
6	0	0	3	1	1	0	0	0	0	0	1	4	CC

Cuadro 3.8: Formato de la base de datos Final (Columnas 1:13 y 6 primeras filas)

Posteriormente se realiza una modificación en los nombres de los SNP, agregando el cromosoma en que se encuentran como un número separado por un punto al final del nombre. Por ejemplo el SNP rs55885037.1, se ubica en el cromosoma 1.

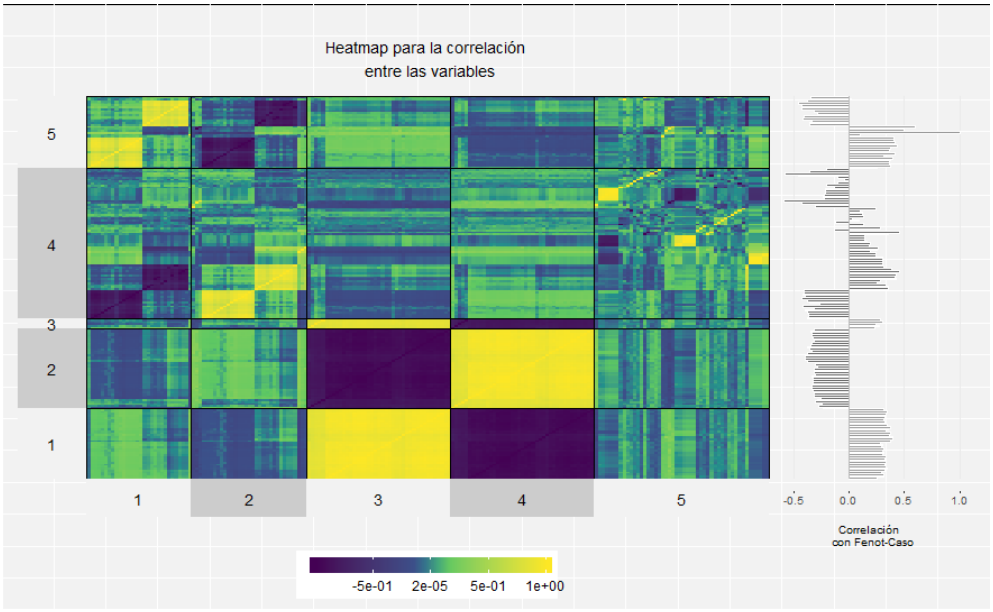


Figura 3.8: Se observan bajas correlaciones con el fenotipo, a excepción del extremo superior del grupo 4.

Con el fin de utilizar la base para la generación de diferentes modelos se decide considerar las diferentes categóricas de cada variable como categorías *dummy*, lo que da como resultado 204 variables dicotómicas. Para esta base resultante se analizan las correlaciones mediante un mapa de calor presente en la Figura 3.8. Se observar agrupa-

ciones en su mayoría con correlaciones menores a 0.5. Los genotipos interrelacionados en los grupos 1 y 3 corresponden a variables dicotómicas y por tanto mutuamente excluyentes. Por su parte las variable que presentan mayor correlación con el fenotipo son particularmente de la *Base 1*. A excepción de estas agrupaciones se podría suponer independencia; sin embargo este hecho se discutirá en el Capítulo 5.

Posición	Variable	Asociación con Fenot Caso	Posición	Variable	Asociación con Fenot Caso
1	Nage18_29	-4.57E-01	181	rs115829688.1AG	1.63E-01
2	PolNo	-2.83E-01	182	rs74455361.1TC	1.64E-01
3	Nage30_41	-2.81E-01	183	rs73323468.14GA	1.65E-01
4	FWCCNo	-2.53E-01	184	rs192567137.5TC	1.67E-01
5	ObeNo	-2.39E-01	185	rs55885037.1TC	1.69E-01
6	DiabNo	-2.37E-01	186	rs1803767.19CT	1.74E-01
7	rs76284034.1GG	-1.95E-01	187	rs115132597.12TC	1.75E-01
8	rs115246312.18GG	-1.79E-01	188	rs118184226.12AG	1.78E-01
9	rs118184226.12GG	-1.78E-01	189	rs115246312.18AG	1.79E-01
10	rs115132597.12CC	-1.75E-01	190	rs76284034.1AG	1.95E-01
11	rs1803767.19TT	-1.74E-01	191	DiabSi	2.37E-01
12	Alc75_150	-1.70E-01	192	ObeSi	2.39E-01
13	rs55885037.1CC	-1.69E-01	193	FWCCSi	2.53E-01
14	rs192567137.5CC	-1.67E-01	194	PolSi	2.83E-01
15	rs73323468.14AA	-1.65E-01	195	Nage55_68	4.89E-01

Cuadro 3.9: Variables representativas del mapa de calor. Se muestran las 15 variables más asociadas a los Fenotipos Caso - Control.

3.3.1. Ajuste de un modelo

Iniesta et al. [24] hace hincapié en la importancia de ajustar un modelo adecuado al análisis de SNP. En particular el artículo en mención recomienda el uso de *haplo.Stats23* como ayuda para la realización de los análisis asociados a los SNP; sin embargo para el estudio de las características estadísticas de los SNP se utilizó el paquete *SNPassoc* [22], - el cual incluye a *haplo.Stats23* - que permite establecer un modelo a partir de gráficos tipo Manhattan - Gráficos de p valores-.

- Modelo codominante: Cada genotipo proporciona un riesgo diferente no aditivo
- Modelo dominante: Supone que una única copia aléllica es suficiente para modificar el riego y por tanto portadores homocigotos y heterocigotos tienen el mismo riesgo.
- Modelo recesivo: Supone que son necesarias dos copias del polimorfismo para que se presente el fenotipo.

- Modelo aditivo: Supone que cada copia alellica del SNP modifica el riesgo en una cantidad aditiva

Como ejemplo, considere que un SNP tiene un polimorfismo "C" que modifica el riesgo de padecimiento. Bajo los modelos anteriormente expuestos un individuo homocigoto tiene un genotipo "CC", mientras que un heterocigoto ""CT" o "TC" -Para nuestros efectos equivalentes- la codificación para cada modelo corresponde a:

	Codominante		Dominante	Recesivo	Aditivo
Genotipo	He	Va	Do	Re	Ad
TT	0	0	0	0	0
TC	0	1	1	0	1
CC	1	0	1	1	2

Cuadro 3.10: Codificación modelos de herencia para un polimorfismo en un *locus* bialelico T>C.

En la Figura 3.9 el modelo codominante - dominante - log aditivo muestra significancias que coinciden con las indicadas por el equipo de Uruguay; esto a partir de la importancia de sus p - valores. Para un modelo codominante, de acuerdo con [24] se tiene la forma.

$$\log \frac{p}{1-p} = \alpha + \beta_1 He + \beta_2 Va + \gamma Z \tag{3.2}$$

En este modelo la variable **Z** representa los componentes no genéticos independientes, **He** a los genotipos heterocigotos, **Va** a los genotipos homocigotos variable o de menor frecuencia y α una constante a determinar. Note que esta codificación es equivalente a considerar *dummy* cada uno de los genotipos para un SNP. como se muestra en el Cuadro 3.10.

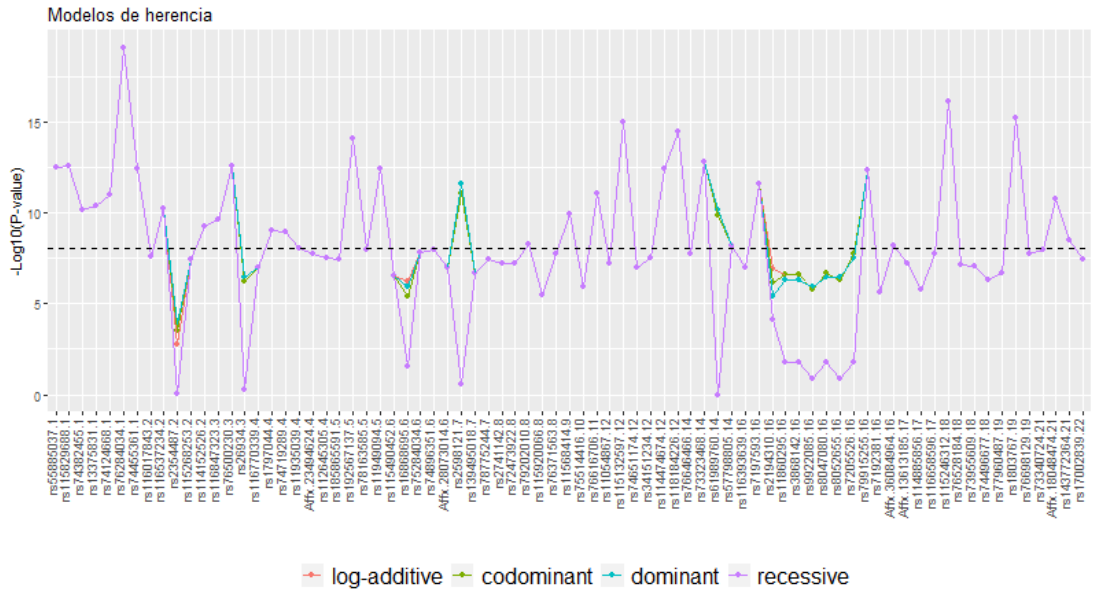


Figura 3.9: Comparativa de la información suministrada por los diferentes modelos.

Equilibrio de Hardy Weinberg (HW)

El equilibrio de HW es importante pues corresponde a una prueba estadística de tipo χ^2 en la que se evalúa la distribución genotípica de los SNP en una población. La Ley de HW establece que salvo mutaciones, las proporciones genéticas en la población permanecen en equilibrio.

$$P(aa) + 2P(aA) + P(AA) = 1$$

Hay que notar, que la ley de HW no indica que el desequilibrio esté asociado a una enfermedad como tal, lo que hace es dar valores de referencia para observar la incidencia en la población y ayudar a identificar potenciales SNP asociados a un fenotipo. La función *WGassociation* suministra esta característica, entre otras de interés.

En particular el Cuadro 3.11 muestra algunos SNP que presentan desequilibrio marcado, como lo son el 37 - rs2194310.16 con una frecuencia de 69 /100 y el 38 - rs2354487.2 con una frecuencia de 80 / 100 lo cual podría corresponder a SNP que están mutando en la población.

	snp	allelos	freq	HW	OR	L95	U95	p.value
5	Affx.36084964.16	C/T	9.8E+01	1	5.7E+00	2.9E+00	1.1E+01	6.2E-09
15	rs11568414.9	C/A	9.8E+01	1	1.0E+01	4.1E+00	2.6E+01	1.1E-10
20	rs116537234.2	G/A	9.8E+01	1	8.3E+00	3.8E+00	1.8E+01	5.8E-11
23	rs116847323.3	C/T	9.8E+01	1	7.3E+00	3.4E+00	1.5E+01	2.5E-10
26	rs11935039.4	G/A	9.8E+01	1	4.5E+00	2.5E+00	8.0E+00	9.0E-09
30	rs143772364.21	G/A	9.8E+01	1	5.2E+00	2.8E+00	9.8E+00	3.5E-09
45	rs57798805.14	C/A	9.8E+01	1	5.1E+00	2.7E+00	9.7E+00	6.2E-09
59	rs74719289.4	G/A	9.8E+01	1	5.2E+00	2.8E+00	9.6E+00	1.1E-09
73	rs79202010.8	C/T	9.8E+01	1	6.5E+00	3.1E+00	1.4E+01	5.5E-09
14	rs115490452.6	C/T	9.8E+01	6.2E-01	3.5E+00	2.1E+00	5.8E+00	2.8E-07
37	rs2194310.16	C/T	6.9E+01	4.0E-01	1.5E+00	1.3E+00	1.7E+00	1.3E-07
38	rs2354487.2	T/C	8.0E+01	5.9E-02	1.3E+00	1.1E+00	1.5E+00	1.6E-03
42	rs34151234.12	G/A	9.6E+01	2.7E-01	3.0E+00	2.0E+00	4.5E+00	3.1E-08
57	rs74496677.18	C/T	9.7E+01	6.3E-01	3.1E+00	1.9E+00	5.0E+00	5.4E-07

Cuadro 3.11: Se muestran las OR significativas de acuerdo al estudio de asociación. Se filtran los datos por $p - value < 10^{-8}$ en el primer caso y luego por HW<1 en el segundo; en ambos se considera OR>1.

Como se observa, el programa nos da de forma directa la penetrancia de los SNP; determinando algunos con OR de valores interesantes; mismos que se verificaran en los análisis posteriores.

Capítulo 4

Análisis de Correspondencia

En esta sección se realiza un Análisis de Correspondencia Múltiple (MCA) a las bases de datos. La naturaleza nominal de las variables y los objetivos de asociación entre las mismas, así como la comparativa con otros análisis con fines similares hacen considerar MCA como la orientación más adecuada. Los estudios se apoyan en *FactoMineR* y *factoextra* paquetes de *R* .

Correspondiente a una técnica de reducción de dimensión como el análisis por componentes principales (PCA), MCA se caracteriza por estar centrado en la incorporación variables cualitativas y sus interrelaciones, las coordenadas de cada punto en el plano corresponden a pruebas χ^2 generadas a partir de las frecuencias entre los cruces de variables, estas se interpretan como indicador de influencia a partir de la distancia respecto al origen del eje coordenado. A mayor distancia del centro por parte de una variable mayor influencia y poder discriminatorio representa. Información sobre el tema se puede encontrar en [8, 23].

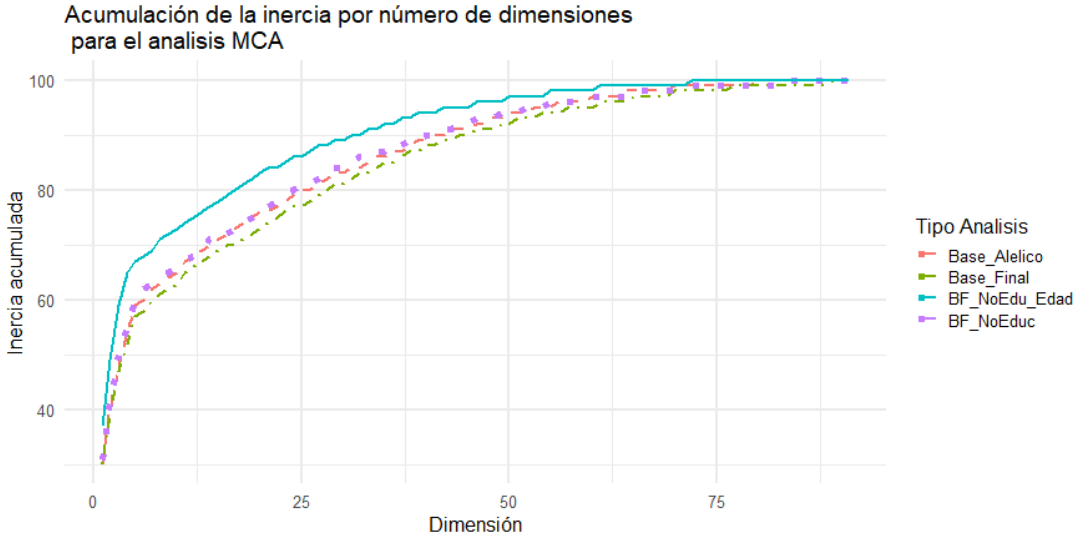
Como cualquiera de los métodos de reducción dimensional, MCA intenta describir la variabilidad de los datos -llamada en este modelo inercia- en una dimensión menor a la original, apoyado en el uso de tablas de contingencia. La matriz de frecuencias resultantes (**Matriz de Burt**) es utilizada para realizar los análisis de forma similar a la matriz de correlaciones utilizada en PCA.

4.1. Análisis de Correspondencia Múltiple (MCA)

El paquete MCA de la librería *FactoMineR* [32] es seleccionado para realizar los cálculos, dada su versatilidad en la gráficación, etiquetado y presentación de los resultados, así como la multitud de opciones que calcula de forma predeterminada (varianza, coordenadas, explicación y representación de las variables por eje, entre otros).

Inicialmente se realiza un estudio de la explicación de la inercia para determinar el número de dimensiones a utilizar, en este se considera la *Base Final* -la cual tiene 204 variables -y una base alélica [1] la cual permite concluir que para el análisis MCA el estudio orientado en genotipos a portan mayor información respecto al orientado a los alelos.

Los resultados de la Figura 4.1 permiten observar que la *Base Final* aumenta su inercia para un mismo número de dimensiones al eliminar los niveles educativos *"Ed"* -BF_NoEduc- y los grupos de edad *"Nage"* -BF_NoEd_Edad- obteniendo una mejora de entre 3 % a 4 % , este hecho no se mantiene con otras variables en las cuales su eliminación conduce a una reducción de la inercia.



	dim 1	dim 2	dim 3	dim 4	dim 5	dim 6	dim 7	dim 8	dim 9	dim 10
Base_Alelico	30.00	40.00	48.00	54.00	59.00	60.00	62.00	63.00	64.00	65.00
Base_Final	30.00	39.00	47.00	52.00	57.00	58.00	60.00	61.00	62.00	63.00
BF_NoEduc	31.00	41.00	50.00	55.00	60.00	62.00	63.00	64.00	65.00	66.00
BF_NoEdu_Edad	33.00	43.00	52.00	57.00	62.00	64.00	65.00	66.00	67.00	69.00

Figura 4.1: Acumulación de la inercia por tipo de análisis y dimensión.

La Figura 4.2 presenta una proyección de los individuos y de las variables en las primeras dos dimensiones acumulando un 43 % de la inercia- lo que contrasta con el 16 % explicado por las dos primeras componentes del análisis PCA indicado en la Figura 3.6-. En triángulos azules se muestra a las variables, mientras que los individuos son indicados en puntos magenta (Controles) y naranja (Casos). En MCA una variable o individuo es significativo en función de su distancia al centro, estas distancias corres-

[1] En esta no se combinaron los alelos, generando una base de 267 variables.

ponden al estadístico χ^2, considerando como grados de libertad al número de individuos o variables utilizados.

De esta manera se interpreta como datos afines a aquellos que se encuentren "cercanos" entre sí, los cuales se esperaría compartan características similares.

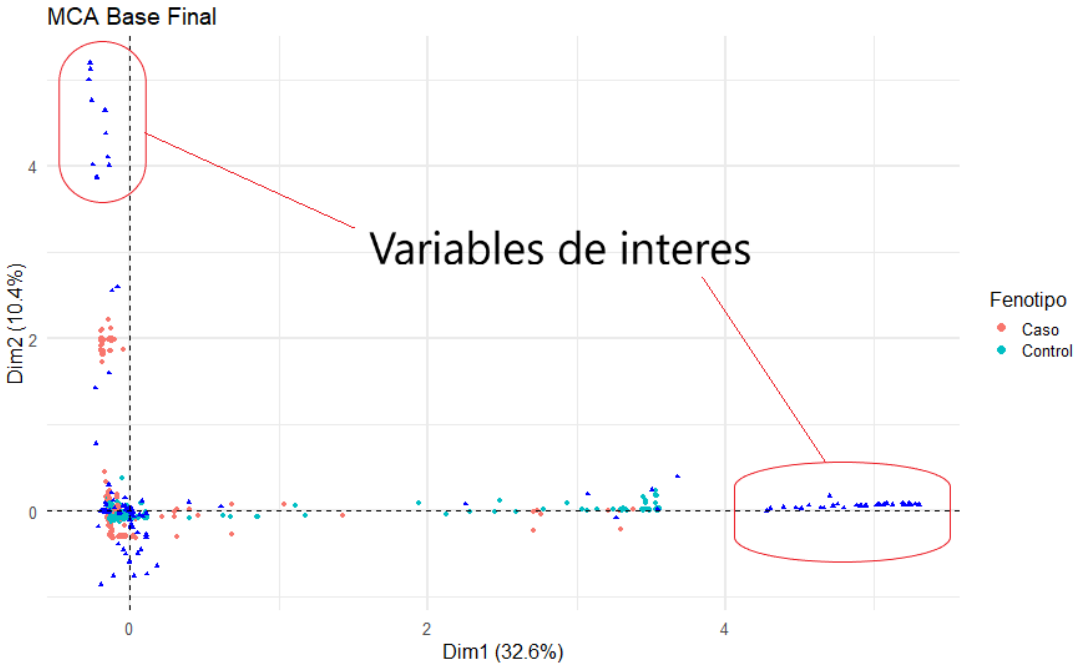


Figura 4.2: Gráfico de correspondencias múltiples para el total de variables e individuos.

De lo anterior, son de particular interés los individuos y variables ubicados en los extremos superiores y derechos de la gráfica, al corresponder a los datos más discriminadores. En contraste, los datos centrales no tienen un efecto relevante representando a variables confusoras.

El detalle de la Figura 4.3 muestra que las variables poco significativas corresponde a variables no genómicas, las cuales no son correctamente representadas en las primeras dimensiones -En particular en las primeras 5-, además la poca variabilidad a lo largo del eje X indicada a partir de la alineación vertical de las observaciones es muestra de una mayor similitud en las características de las variables asociadas al eje X, las cuales como se verá corresponden a *"controles"*.

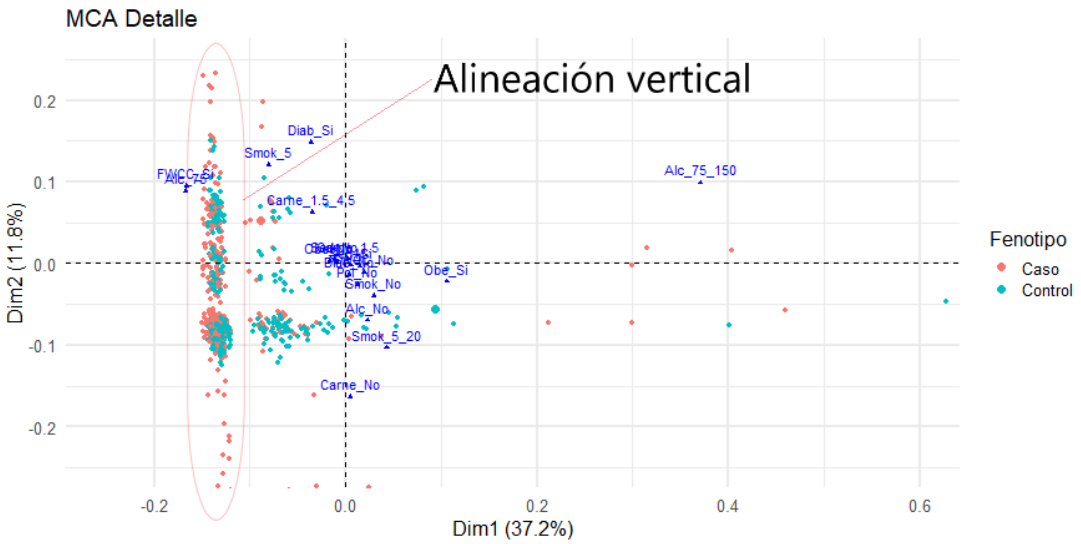


Figura 4.3: Detalle de variables poco representativas para el análisis de MCA

Con el fin de recopilar sólo los predictores que aporten mayor información, se considerará significativa a una variable en relación a su aporte medio, definiendo éste como el promedio de la inercia explicada por las variables más contributivas de cada dimensión [29, pág 59], esto quiere decir que aquellas variables con aporte menor a una contribución umbral de 5 % no son consideradas en el análisis, bajo este esquema las variables eliminadas se muestran en el Cuadro 4.1.

Smok_20	Alc_150	Carne_4.5	rs2354487.2_CC
rs26934.3_AA	rs16868695.6_TT	rs2598121.7_CC	rs11860295.16_TT
rs3868142.16_AA	rs9922085.16_CC	rs8047080.16_GG	rs8052655.16_AA
			rs7205526.16_TT

Cuadro 4.1: Variables eliminadas del análisis MCA, la contribución de ninguna ellas no alcanza el 5 % para las 10 dimensiones consideradas.

La ventilación de las variables supone una mejora de entre 7 % y 10 % en la inercia explicada por cada dimensión manteniendo 177 de los 190 predictores originales. En el Cuadro 4.2 se observa la variación de contribución de la inercia a lo largo de las 10 dimensiones consideradas permite alcanzar un aproximado de 73 % de inercia

	dim 1	dim 2	dim 3	dim 4	dim 5	dim 6	dim 7	dim 8	dim 9	dim 10
Base_Final	30.00	39.00	47.00	52.00	57.00	58.00	60.00	61.00	62.00	63.00
BF umbral (5 %)	37.19	48.99	58.65	64.83	66.55	67.97	69.30	70.53	71.72	72.88

Cuadro 4.2: Comparativa en la contribución de la inercia

Considerando las variables ventiladas, llama la atención que los genotipos corresponden con variables en que los $\mathbf{p-valores}$ son mínimos para el modelo de codificación hereditario recesivo - Figura 3.9- identificando a 10 de los 12 SNP cuyo valor disminuía respecto a los otros modelos, esto supone un punto de interés al representar una estrategia para el descarte de variables poco discriminadoras, lo que corresponde a un proceso diferente para seleccionar modelos de herencia respecto a la indicada en el Capítulo 3.

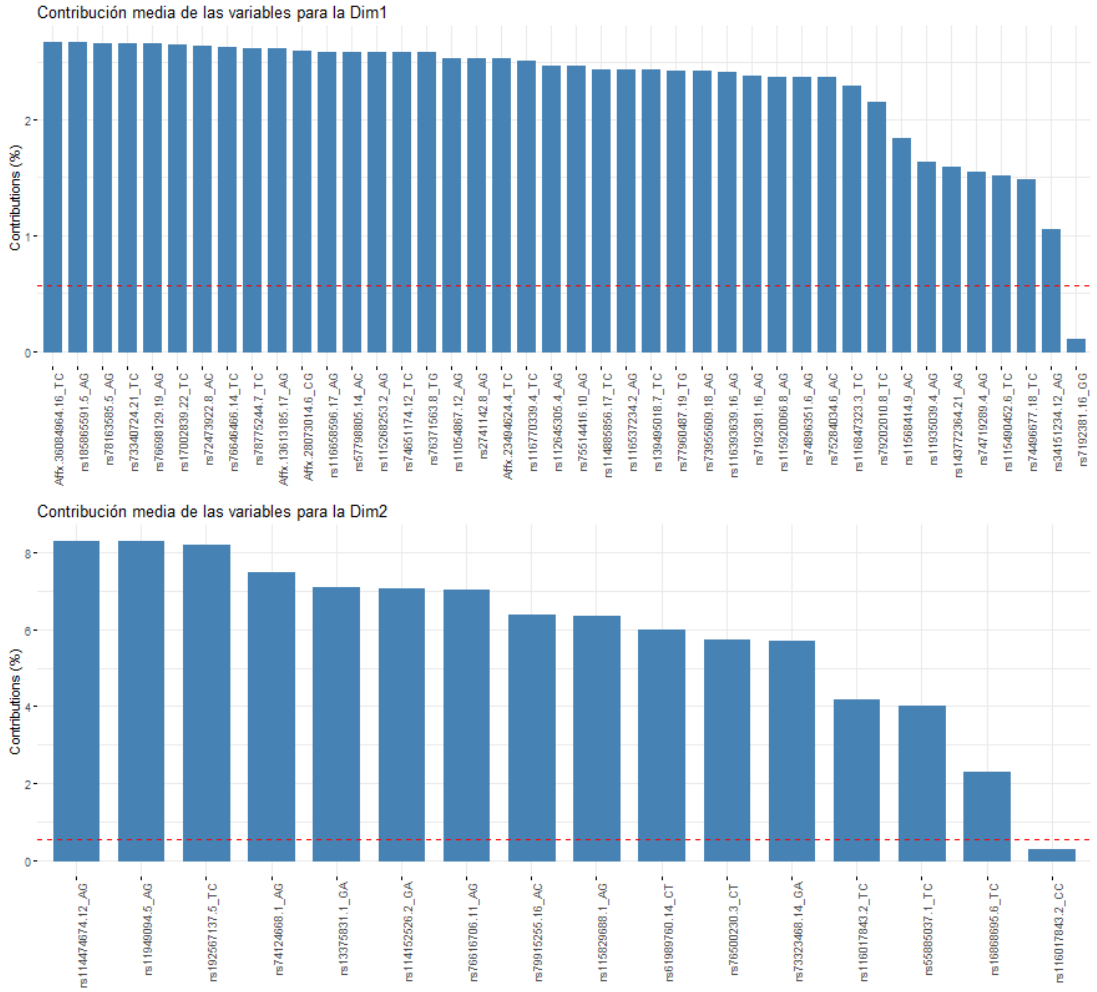


Figura 4.4: Variables contributivas de la dimensión 1 y 2.

Definiendo como significativos a los predictores que se encuentren por encima del aporte medio, se seleccionan 41 variables para la dimensión 1 -indicadas en la Figura 4.4-, 15 para la dimensión 2 y 130 individuos para las dimensiones 1 y 2 en conjunto, incluyendo elipses de concentración al 95 %, se obtiene la Figura 4.5. [1].

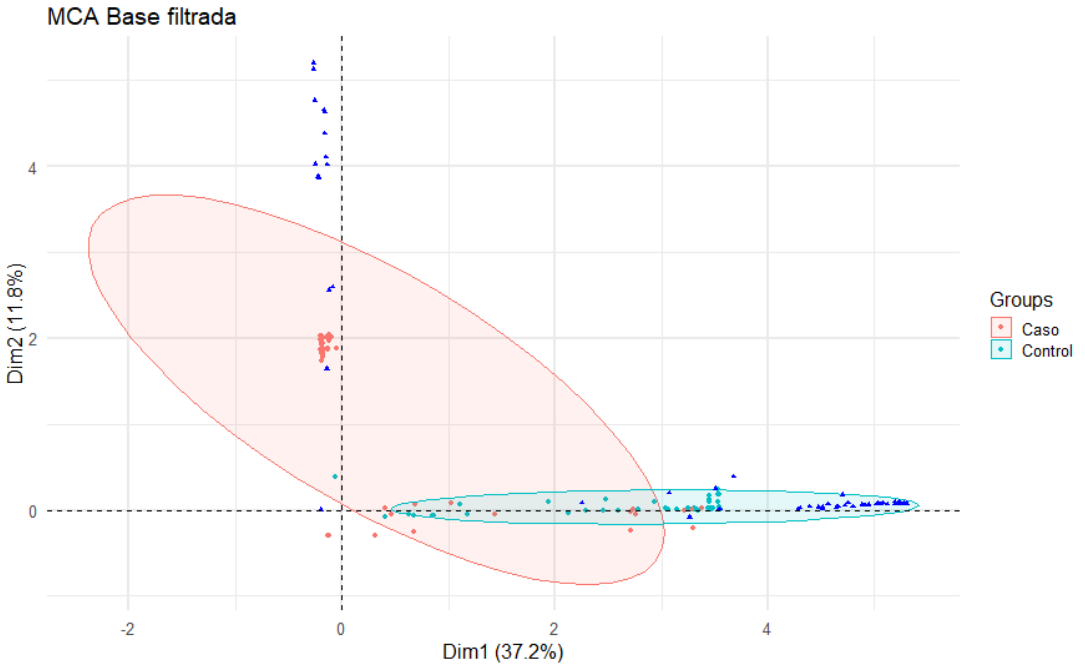


Figura 4.5: MCA de variables e individuos significativos para las dimensiones 1 y 2.

como se aprecia, para las dos primeras dimensiones existe una alta variabilidad de los *"casos"*, mientra que los *"controles"* presentan una concentración a lo largo del eje X, esta concentración permite considerar la hipótesis de que las variables - en este caso genotipos - en el interior de la elipse corresponden a factores protectores para los individuos considerados.

4.2. Interpretación del MCA

La proyección en un mismo plano de las observaciones en conjunto con las variables permite establecer asociaciones. En particular las elipses mostradas en la Figura

[1]Las elipses de concentración se conforman a partir de la variabilidad de los datos, en donde los semiejes son determinados a partir de las varianzas de las coordenadas X, Y de cada grupo - caso, control- el resultado muestra un 95 % de los individuos dando idea de su variabilidad respecto a cada eje [29]

4.5 ofrecen noción sobre comportamiento de los datos, en los cuales se observa una asociación con cada uno de los ejes, siendo el eje X predominantemente abarcado por *"Controles"*.

4.2.1. Asociaciones en los 2 primeros ejes

Las 37 variables dentro del contorno de la elipse extendida a lo largo del eje X comparten características similares -esto bajo las relaciones χ^2- por lo que resulta natural el pensar estén asociadas a factores protectores contra el CCR al observar todos los genotipos en el grupo 2 del mapa de calor del Capítulo 3, los cuales estaban asociados negativamente con los *Casos*.

1	Affx.13613185.17AG	rs115920066.8AG	rs185865591.5AG	rs74719289.4AG
2	Affx.23494624.4TC	rs116393639.16AG	rs2741142.8AG	rs74896351.6AG
3	Affx.28073014.6CG	rs116537234.2AG	rs34151234.12AG	rs75284034.6AC
4	Affx.36084964.16TC	rs116658596.17AG	rs57798805.14AC	rs75514416.10AG
5	rs11054867.12AG	rs116770339.4TC	rs7192381.16AG	rs76371563.8TG
6	rs112645305.4AG	rs116847323.3TC	rs72473922.8AC	rs76646466.14TC
7	rs114885856.17TC	rs11935039.4AG	rs73340724.21TC	rs76698129.19AG
8	rs115268253.2AG	rs139495018.7TC	rs73955609.18AG	rs77960487.19TG
9	rs115490452.6TC	rs143772364.21AG	rs74496677.18TC	rs78163585.5AG
10	rs11568414.9AC	rs17002839.22TC	rs74651174.12TC	rs78775244.7TC
11	rs79202010.8TC			

Cuadro 4.3: Potenciales genotipos protectores

En la Figura 4.5 la agrupación de individuos *"Caso"* a la altura $\chi^2_y = 2$ corresponde a 53 observaciones que llaman la atención, esto al corresponder a individuos diferenciadores para las dimensiones presentadas. Analizando sus características observamos que el 100 % presenta los 45 genotipos indicados en el Cuadro 4.4 las características que presentan son

- 36 están presentes en el mapa de calor en los grupos 1 y 3 mostrando una correlación positiva con los *"Casos"*.
- 6 forman parte de los 17 genotipos asociados al eje Y -indicados en negrita en el Cuadro 4.4- y pertenecientes al grupo 5 de las filas.
- Los 3 genotipos faltantes forman parte del grupo 4.

De los 45 genotipos causales, 41 son dicotómicos respecto a las variantes protectoras. Luego la independencia de los ejes se refleja en la independencia entre genotipos de donde se concluye que al menos los 41 genotipos asociados son diferenciadores entre *casos* y *controles*.

1	**rs74124668.1AG**	rs115490452.6CC	rs75514416.10GG	Affx.36084964.16CC
2	**rs116017843.2TC**	rs75284034.6CC	**rs76616706.11AG**	Affx.13613185.17GG
3	rs116537234.2GG	rs74896351.6GG	rs11054867.12GG	rs114885856.17CC
4	rs115268253.2GG	Affx.28073014.6GG	rs74651174.12CC	rs116658596.17GG
5	rs116770339.4CC	rs139495018.7CC	**rs114474674.12AG**	rs115246312.18GG
6	rs11935039.4GG	rs78775244.7CC	rs76646466.14CC	rs76528184.18CC
7	Affx.23494624.4CC	rs2741142.8GG	rs57798805.14CC	rs73955609.18GG
8	rs112645305.4GG	rs72473922.8CC	rs116393639.16GG	rs77960487.19GG
9	rs185865591.5GG	rs79202010.8CC	rs7192381.16GG	rs1803767.19TT
10	**rs192567137.5TC**	rs115920066.8GG	rs17002839.22CC	rs76698129.19GG
11	rs78163585.5GG	rs76371563.8GG	rs11568414.9CC	rs73340724.21CC
12	**rs11949094.5AG**			

Cuadro 4.4: Genotipos asociados a 53 casos diferenciados por MCA

4.2.2. Análisis sobre las variables

La inercia concentrada en la primera dimensión -Figura 4.6- implica una explicación puntal del fenómeno de un 37 % siendo el eje X de la dimensión 1 un fuerte caracterizador de los individuos *control*.

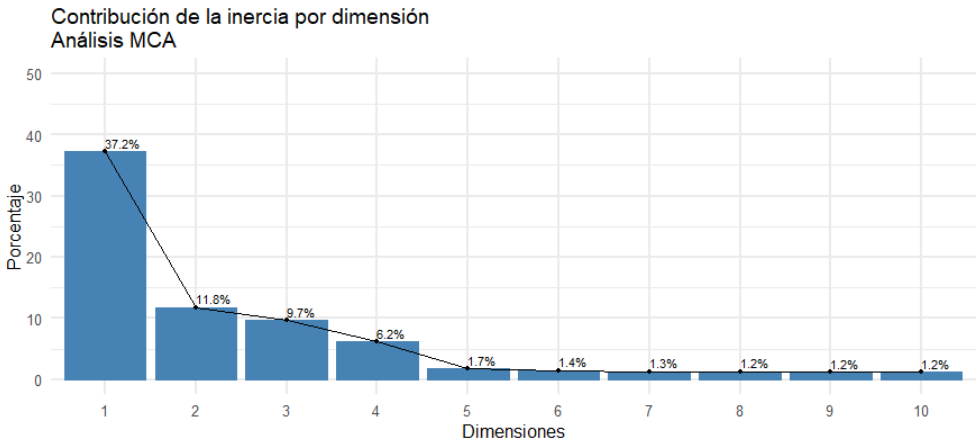


Figura 4.6: La contribución de la inercia disminuye rápidamente a lo largo de las 10 dimensiones consideradas. Las primeras 4 abarcan un 65 % de la explicación de los datos

Los 4 primeros ejes aglomeran un 65 % de la información caracterizados por variables genómicas, hecho que se puede observar directamente en la Figura 4.8 la que permite comparar a todas la dimensiones en conjunto, así como las variables más representativas en cada dimensión. El tono de azul y el tamaño del punto indican la contribución de la variable en la dimensión considerada.

Se observa que la contribución de las variables genotípicas es principalmente puntual y de importancia, al presentarse en la mayoría de los casos solo una vez a lo largo de las primeras cuatro dimensiones; En contraste las variantes ambientales tienen un efecto laxo y distribuido a lo largo de las dimensiones 5 a 10. A lo que cabe la pregunta ¿qué tanto aportan las variables ambientales en conjunto?. Para responder a lo anterior hay que considerar el aporte de cada variable por dimensión, así como la su importancia -inercia que explica-, lo cual se resume en la ecuación (4.1).

$$Explic = \sum_{i,j} contrib\_variable_i \times inercia\_dimension_j \tag{4.1}$$

el cálculo aplicado a las 177 variables permite obtener una distribución donde se distinguen 3 grupos, la caracterización de estos se observa en la Figure 4.7 en donde el grupo de mayor cantidad de variables es el que menor contribución aporta.

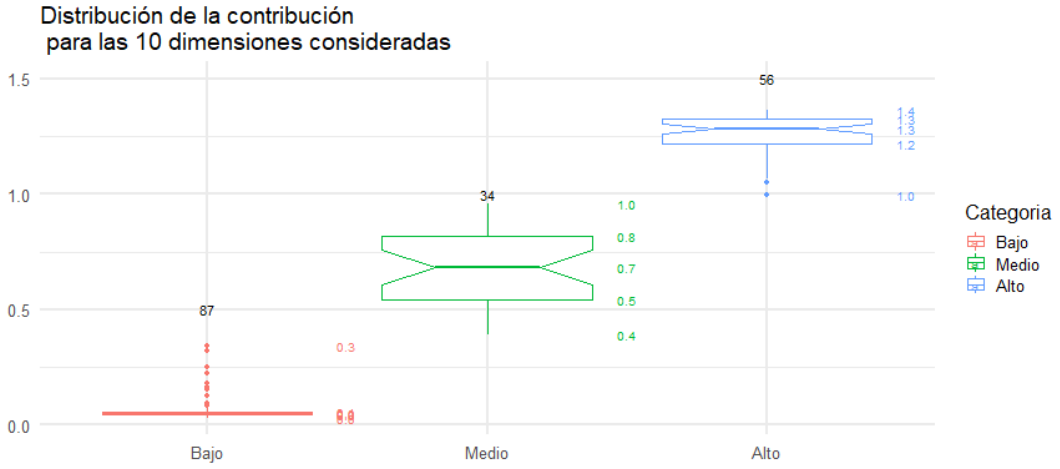


Figura 4.7: Los datos de contribución de las 177 variables permiten observar la presencia de tres agrupaciones principales

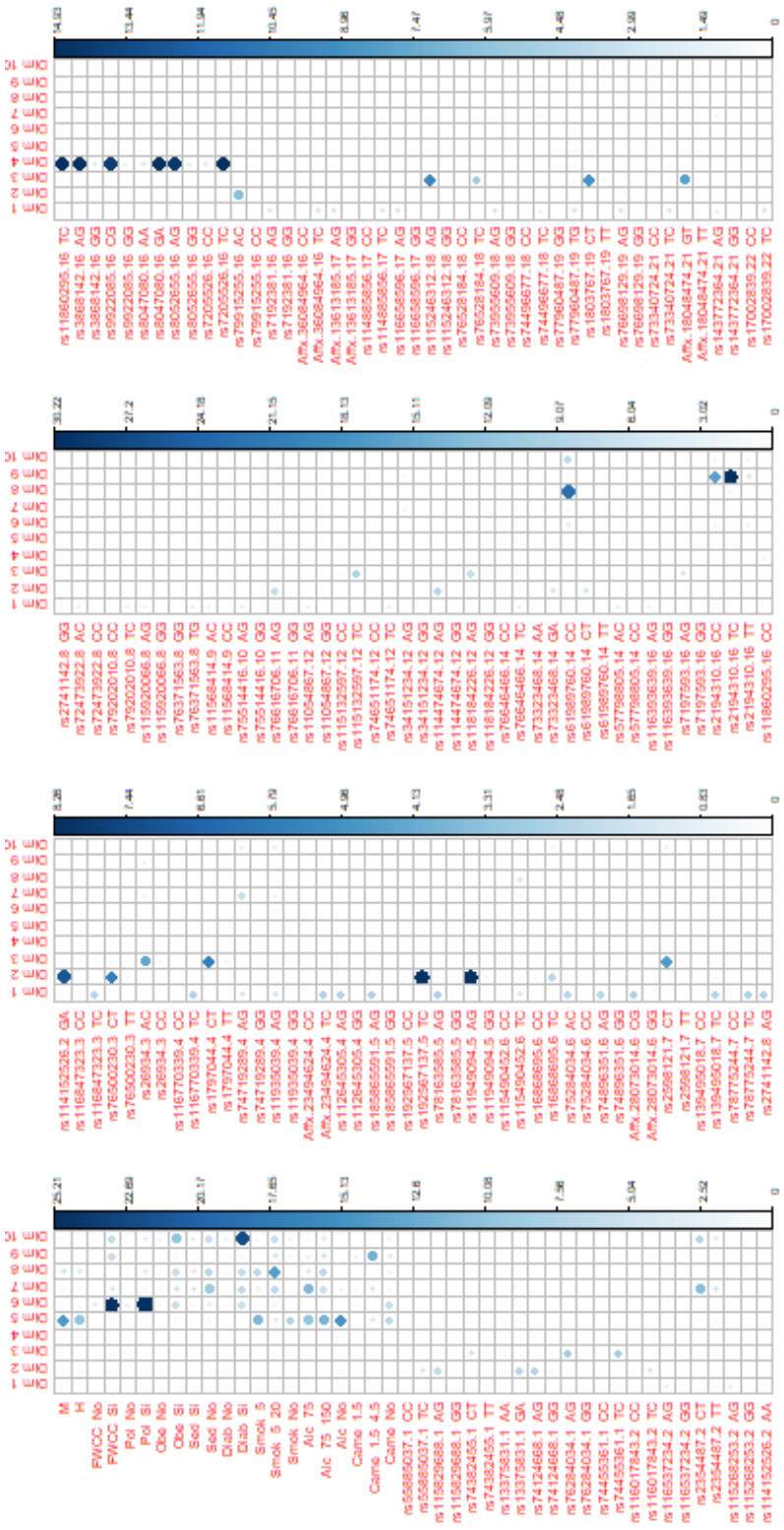


Figura 4.8: Contribución de las variables a la representación de los ejes

En la Figura 4.7 se muestran gráficos de caja para cada grupo, la ranura intermedia corresponde a la mediana y el número ubicado en el extremo superior a la cantidad de datos que engloba. Las características de cada grupo son:

- Grupo de alta contribución (*"Alto"*). Con una contribución media de 1.3 % engloba a 56 genotipos, 34 de los cuales están presentes en la dimensión 1 como protectores, 10 en la dimensión 2 -percusores y los otros 12 se encuentran distribuidos a lo largo de las dimensiones 3 y 4. Explica un 71 % de la inercia total.
- Grupo de media contribución (*"Medio"*). Su contribución es en promedio 0.7 % lo que corresponde poco más del 50 % del grupo superior. Engloba a 33 variables, 13 de las cuales son no genotípicas (PolSi, FWCCSi, DiabSi, ObeSi, Smok5_20, Smok5, M, CarneNo, Carne1.5_4.5, AlcNo, Alc75, Alc75_150, SedNo), 12 de ellas ubicadas en la dimensión 5. Es el grupo de mayor variabilidad, de las 20 variables genotípicas 7 están presentes en la dimensión 1 -serian protectoras -, 4 en la dimensión 2, 7 en la dimensión 3. repitiendo en algunos caso su presencia en otras dimensiones. Un hecho de importancia para estas variables es que el conjunto de SNP es independiente del conjunto de SNP de las variables de contribución *"Alta"*, por lo que se podría pensar que su efecto es independiente del efecto generado por el grupo anterior. Describe un 23 % de la inercia.
- Grupo de baja contribución (*"Baja"*). Su contribución es apenas perceptible con un promedio de 0.08 %. contiene a 8 variables no genotípicas (DiabNo, FWCCNo, H, ObeNo, PolNo, SedSi, SmokNo, Carne1.5) y 80 genotípicas de las cuales 3 se presentan la dimensión 2 y ninguna en la dimensión 1. Tiene muy poca variabilidad, con 16 valores atípicos. En conjunto explica un 6 % de la inercia

La caracterización de cada grupo permite observar que conforme disminuye la contribución global de una variable, aumenta su efecto a lo largo de las diferentes dimensiones -ver Figura 4.8- . En el grupo *"Alto"* los genotipos distribuidos en las primeras 4 dimensiones no muestran relación con las otras de donde se plantea la hipótesis de los SNP involucrados son poco afectados por las variables ambientales o en su defecto los mecanismos reguladores y reparadores del ADN -mencionados en el capítulo 3- no han sido afectados.

El grupo *"Medio"* parece corresponder a grupos de SNP y variables interrelacionadas, no solo por estar presentes en una misma dimensión, si no por la dispersión de su efecto tanto para las variantes genotípicas como epigenéticas, siendo así se puede postular la hipótesis de que los grupos de SNP de este conjunto son de importancia en el diagnostico del CCR dada su variabilidad. Finalmente, el grupo *"Bajo"* presenta multiples interrelaciones, al igual que el grupo medio. Sin embargo parece ser que el alcance de la afectación que estas modificaciones genera es limitado.

En general lo anterior lo podemos resumir en que el grupo *"Alto"* es importante y alto nivel discriminador -entre casos y controles- el grupo *"Medio"* es un poco impor-

tante con un nivel medio de discriminación, mientra que el grupo *"Bajo"* en general no agrega importancia y tiene un bajo nivel de discriminación.

Como referencia en el Cuadro 4.5 se muestran las 15 variables que más contribuyen en cada grupo y su porcentaje de *"explicación general"* a la inercia para las 10 dimensiones consideradas en MCA.

	N. Alto	%	N. Medio	%	N.Bajo	%
1	Affx.36084964.16_TC	1.37	rs11568414.9_AC	0.96	rs16868695.6_TC	0.38
2	rs185865591.5_AG	1.37	rs118184226.12_AG	0.96	rs2194310.16_CC	0.33
3	rs73340724.21_TC	1.36	rs76500230.3_CT	0.95	H	0.32
4	rs114474674.12_AG	1.36	rs73323468.14_GA	0.94	rs2354487.2_CT	0.28
5	rs11949094.5_AG	1.36	rs11935039.4_AG	0.86	Smok_No	0.25
6	rs76698129.19_AG	1.36	rs143772364.21_AG	0.84	Sed_Si	0.24
7	rs78163585.5_AG	1.36	rs74719289.4_AG	0.84	rs11860295.16_CC	0.16
8	rs17002839.22_TC	1.36	rs115490452.6_TC	0.83	rs8047080.16_AA	0.16
9	rs115246312.18_AG	1.35	Pol_Si	0.81	rs3868142.16_GG	0.16
10	rs72473922.8_AC	1.35	rs1797044.4_CT	0.79	rs7205526.16_CC	0.15
11	rs192567137.5_TC	1.35	rs74496677.18_TC	0.78	rs2194310.16_TT	0.14
12	rs76646466.14_TC	1.34	FWCC_Si	0.77	rs2354487.2_TT	0.13
13	rs78775244.7_TC	1.34	rs7197593.16_AG	0.75	rs9922085.16_GG	0.13
14	Affx.13613185.17_AG	1.33	rs76528184.18_TC	0.74	rs8052655.16_GG	0.12
15	rs57798805.14_AC	1.33	rs74382455.1_CT	0.71	Obe_No	0.10

Cuadro 4.5: Las variables genotípicas de cada grupo, muestran tendencia a la agrupación con variables de la *Base 1* con forme se reduce su participación

Note que en general, en el presente capítulo no solo hemos determinado SNP diferenciadores y sus agrupaciones, si no que además asociaciones entre genotipos protectores -Figura 4.4- y percusores del CCR -Cuadro 4.4 - con las variantes epigenéticas, así como porcentajes de participación en su explicación siendo en el caso de las variables genéticas un mínimo de 70 %; se observo además relaciones entre el parámetro de ventilación y el descarte de variables realizado por la codificación de un modelo regresivo, hecho que pone de manifiesto una utilidad práctica.

Capítulo 5

Diagnóstico mediante métodos de Ciencia de Datos

En el Capítulo 4 se determinaron relaciones de interés entre genotipos protectores así como precursores, se crearon grupos de variables que contribuían en diferentes grados a la explicación de la inercia -equivalente en MCA a la varianza- se observo que variables epigenéticas como *"Smok_20"*,*"Alc_150"* y *"Carne_4.5"* tienen una poca variabilidad al depender de pocas observaciones, lo que en conjunto con las asociaciones del observadas en el Capítulo 3 sugieren una reagrupación. Se realizaron hipótesis sobre la importancia de diferentes grupos de variables, las cuales serán exploradas.

Los análisis están basados en las *Bases* descritas en el Capítulo 3 y mediante los algoritmos explicados en el Capítulo 2, utilizando la paquetería *Caret* [31] de *R*. La validación de los resultados se realizará siguiendo los criterios AIC, Devianza, ROC indicados en la Sección 2.4.

Cada modelo tiene una configuración de remuestreo con *validación cruzada - 10 folds - con 5 repeticiones*, esto quiere decir que cada 10 modelos se realiza una nueva selección de observaciones -10 folds- para volver a repetir el algoritmo.[1], lo tiene como fin minimizar la influencia que tiene la separación de los conjunto de entrenamiento y prueba al tiempo que reduce el sesgo.

Al trabajar con varios algoritmos se requiere de un conjunto de prueba para poder compararlos. Para esto se toma una muestra aleatoria del 169 observaciones -aproximadamente 10 % del total 1707- para validar los diferentes modelos propuestos. El porcentaje surge de considerar

- $N = 1707$ Número de observaciones de la *Base Final*

[1] Referencias adicionales sobre el método de validación cruzada en [12, 18].

- $Z = 2{,}57$ Nivel de confianza esperado en los resultados- 99 % a partir del número desviaciones estándar de una distribución normal de varianza 1. El porcentaje de confianza corresponde al estándar de la industria medica.
- $p = \frac{879}{1707} = 0{,}51$ La proporción de *Casos* y
- $\epsilon = 0{,}01 = 1\,\%$ Error máximo a tolerar.

En la ecuación (5.1) de proporción de muestras

$$n = \frac{N \times Z^2 p(1-p)}{(N-1) \times \epsilon^2 + Z^2 \times p(1-p)} \tag{5.1}$$

la cual es utilizada para definir el tamaño de muestra de un población que se distribuye normalmente.

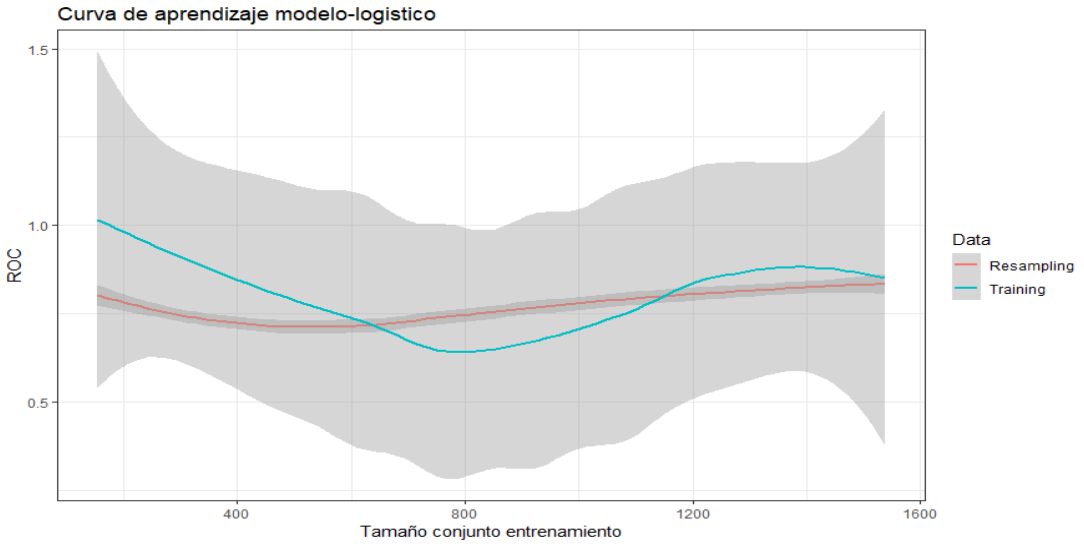
5.1. Modelo logístico

Para la selección de conjunto de entrenamiento y prueba se utiliza una curva de aprendizaje generada a partir de los datos de la *Base Final* con las variables tipo *dummy*, lo que permite abarcar las codificaciones de los modelos de herencia señalados en la sección 3.3.1.

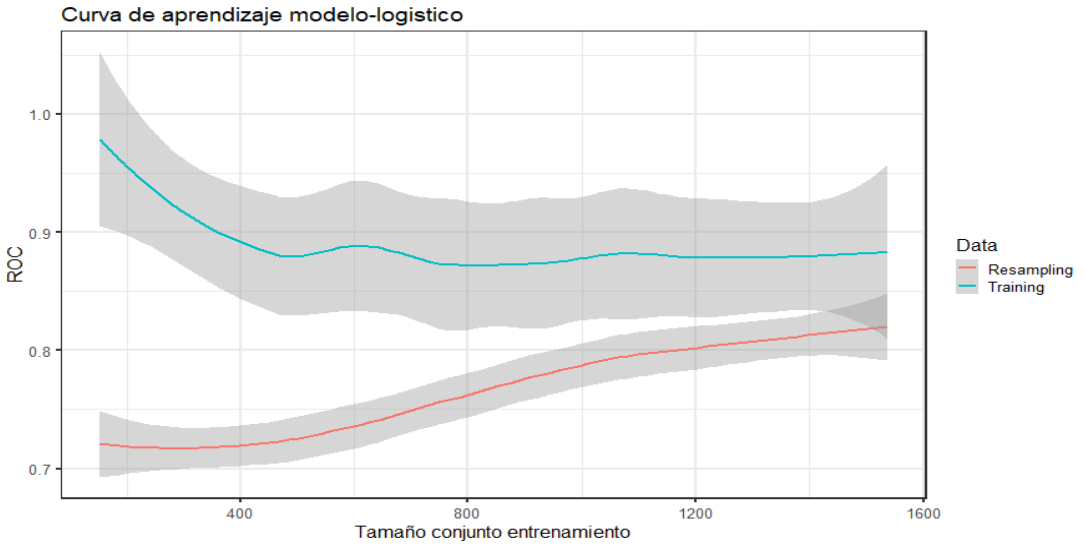
De las variables eliminadas en el análisis MCA Cuadro 4.1 y de las agrupaciones determinadas mediante MDS Figura 3.7, se propone la creación de nuevos predictores a partir de la combinación de variables

- Est_Basico$\sim$ EdPrComp + EdNo
- Est_Medio $\sim$ EdEstpgs + EdEstTec + EdEUniv
- Est_Esp $\sim$ EdSecCom + EdPrep
- SmokSi $\sim$ Smok5 + Smok5_20 + Smok20
- AlcSi $\sim$ Alc75 + Alc75_150 + Alc150
- CarneSi $\sim$ Carne1.5 + Carne1.5_4.5 + Carne4.5

Las curvas de la Figura 5.1 son generadas al entrenar 50 modelos logísticos, la alta variabilidad de la curva de entrenamiento -trainning- mostrada por los intervalos de confianza al 95 % -bandas en gris- y su falta de estabilidad 5.1(a) caracterizada por la variación en la monotonía, son indicadores de sobreajuste o la presencia de variables confusoras.



(a) ***Con combinación de variables***



(b) ***Sin agrupación de variables***

Figura 5.1: La falta de estabilidad de la curva de entrenamiento es indicador de sobre-ajuste o del efecto de variables confusoras

La curva de remuestreo -resampling- corresponde al promedio de los resultados de evaluación para cada uno de los 5 modelos generados, en cada uno de los 10 folds.

Al utilizar las técnicas de selección de variables indicadas en la Sección 2.1.1 se ob-

serva que no se genera una modificación perceptible en los indicadores de selección AIC, DEV, ni en se da una reducción del sobreajuste, evidenciando que el mismo es generado a partir de la combinación de variables.

Como referencia se considera el modelo logit generado por la selección obtenidas mediante el método de *LASSO*, de los 38 estimadores obtenidos se muestran los 15 más significativos en orden creciente respecto a su $\mathbf{p-valor}$

En el Cuadro 5.1 -y los siguientes- el "estimador" corresponde al coeficiente del modelo de regresión logístico, "Std.Error" a la desviación estándar del estimador anterior. "z.value" o el z - valor es el número de desviaciones estándar -valor anterior- a las que se encuentra el estimador del cero - en general de la media- e indica que tan probable es que el estimador sea cero, esta probabilidad es dada por el "p.value" o p-valor.

	variable	estimador	Std.Error	z.value	p.value	OR	sd.OR	LOR	UOR
1	Nage18_29	3.59E+00	5.11E-01	7.03E+00	2.09E-12	2.76E-02	6.00E-01	-1.15E+00	1.20E+00
2	Nage30_41	1.88E+00	2.89E-01	6.49E+00	8.73E-11	1.53E-01	7.49E-01	-1.31E+00	1.62E+00
3	Nage55_68	-2.15E+00	3.43E-01	-6.27E+00	3.71E-10	8.56E+00	7.10E-01	7.17E+00	9.95E+00
4	FWCCNo	3.10E+00	5.22E-01	5.95E+00	2.72E-09	4.49E-02	5.93E-01	-1.12E+00	1.21E+00
5	ObeNo	1.72E+00	3.68E-01	4.69E+00	2.80E-06	1.78E-01	6.92E-01	-1.18E+00	1.54E+00
6	rs2354487.2CT	1.25E+00	2.68E-01	4.64E+00	3.47E-06	2.88E-01	7.65E-01	-1.21E+00	1.79E+00
7	SedSi	-1.02E+00	2.59E-01	-3.93E+00	8.66E-05	2.77E+00	7.72E-01	1.26E+00	4.28E+00
8	rs7205526.16CC	1.57E+00	4.11E-01	3.83E+00	1.30E-04	2.07E-01	6.63E-01	-1.09E+00	1.51E+00
9	CarneNo	1.65E+00	4.77E-01	3.46E+00	5.41E-04	1.92E-01	6.21E-01	-1.02E+00	1.41E+00
10	Nage69_91	-3.80E+00	1.12E+00	-3.40E+00	6.86E-04	4.46E+01	3.27E-01	4.40E+01	4.53E+01
11	DiabNo	2.03E+00	6.68E-01	3.03E+00	2.42E-03	1.32E-01	5.13E-01	-8.73E-01	1.14E+00
12	Est_Medio	7.62E-01	2.57E-01	2.97E+00	3.01E-03	4.67E-01	7.74E-01	-1.05E+00	1.98E+00
13	rs2194310.16CC	-6.68E-01	2.44E-01	-2.73E+00	6.30E-03	1.95E+00	7.83E-01	4.15E-01	3.48E+00
14	rs2194310.16TT	1.20E+00	4.73E-01	2.54E+00	1.10E-02	3.01E-01	6.23E-01	-9.21E-01	1.52E+00
15	Est_Basico	-9.02E-01	3.56E-01	-2.53E+00	1.14E-02	2.46E+00	7.00E-01	1.09E+00	3.84E+00

Cuadro 5.1: Características modelo logit -LASSO utilizando combinación de variables

Por su parte los OR's son calculados a partir de los estimadores $OR = \exp(-\beta)$ donde β representa al estimador. LOR o limite inferior y UOR o limite superior son considerados a partir del coeficiente $\pm 1{,}96$ el cual representa 2 desviaciones estándar de una distribución normal. ver Ecuación (2.20)

5.1.1. Ajuste de un modelo mediante selección de variables

A partir de los 203 predictores y observando que la curva de entrenamiento de la Figura 5.1(b) alcanza estabilidad a partir de aproximadamente 800 observaciones, se

selecciona el máximo de la curva estable, el cual es generado en 1075 observaciones con un ROC de alrededor de 0.89.

Para el modelo logístico el considerar la selección de variables con la codificación binaria de los SNP, tiene la ventaja de permitir el desarrollo de un modelo híbrido, el cual permite el ajuste más adecuado para cada SNP permitiendo sus interacciones. **La importancia de esta selección motiva analizar las diferentes estrategias por separado**.

- Eliminación de predictores "casi nulos" **NZV**: Considere inicialmente la eliminación de las variables que se manifiestan de forma esporádica en las observaciones -Se utiliza el umbral predeterminado de 5 %- la función *nearZeroVar* de *caret* permite determinar que 134 de las 203 variables presentan tal comportamiento. El modelo resultante tiene 72 predictores, de los cuales 22 no son considerados, esto al tener una correlación perfecta con otras -son variables complementarias a otras por ejemplo en genero Hombre y Mujer-. Luego, los 42 predictores resultantes corresponderán a un modelo en el cual no se presenten problemas de inflación de la varianza, evitando la variabilidad de los coeficientes la cual genera desajuste.

- Eliminación de multicolinealidad **COL**: Similar al problema anterior, se refiere a la combinación lineal de columnas. La combinación lineal, permite observar la representación de uno o varios predictores en otro, siendo afectada la precisión de los resultados por la redundancia de los datos. La función *FindLinealCombos* de *Caret* determina 95 linealidades 87 de la cuales están asociadas a los géneros. El problema de la multicolinealidad es inevitable al trabajar con variables *dummy* dado que para las 89 variables iniciales -antes de considerarlas *dummy*- se van a generar al menos 89 problemas de multicolinealidad mismo que se evita eliminando al menos 1 nivel de cada variable. El modelo resultante tiene 100 predictores, uno de los cuales no es considerado por problemas de poca varianza.

- Eliminación de Correlaciones **COR**: En la Figura 3.5 la cual mostraba las correlaciones tetracoricas entre variables de la *Base 1* se observaron correlaciones débiles". Similar al anterior, da un margen de 95 % de asociación lineal entre las variables, eliminando a 118 predictores de los 85 restantes 8 tienen problemas de poca varianza.

- Reducción del error mediante **LASSO**: Utilizando la función *cv.glmnet* de la librería *glmnet* [19] se determina mediante validación cruzada el parámetro $\lambda = 0{,}004$ el cual reduce a cero 158 coeficientes. Al utilizar en el modelo logístico los 45 coeficientes restantes, se omiten predictores 12 por baja varianza.

- Predicción por **PLS**: El modelo de selección muestra que 5 componentes son adecuados -Figura 5.2 -, sin embargo el remuestreo permite observar que el máximo se obtiene con 10 componentes -Figura 5.2- para estos se utiliza una rotación "varimax" y se filtran las variables no seleccionas en las dimensiones utilizadas

obteniendo 37 variables de las cuales 7 no son utilizadas por multicolinealidad o problemas de varianza.

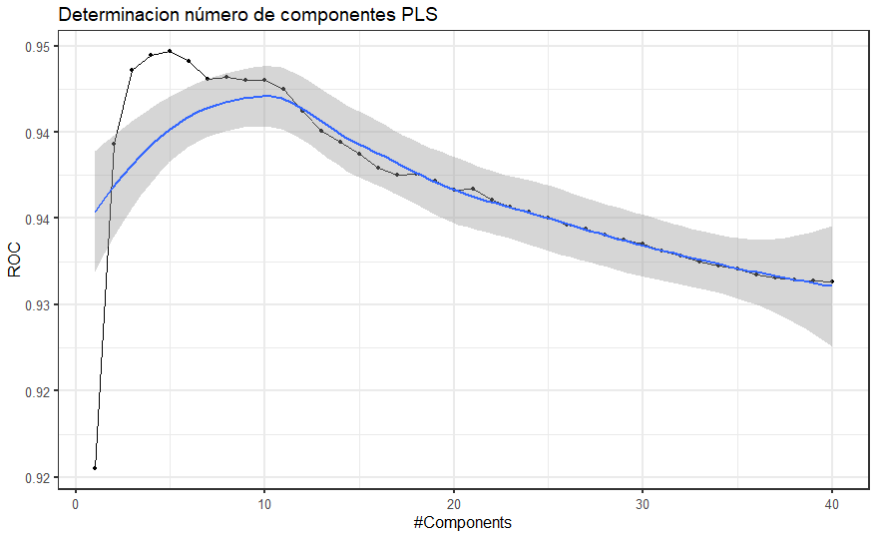


Figura 5.2: El remuestreo permite determinar el número optimo de componentes para PLS

La comparativa entre los diferentes métodos de regularización se presentan en el Cuadro 5.2, el uso de cada modelo depende de los intereses en la retención de variables que se tengan, así como umbral de significancia que se desee utilizar para los **p – valores**.

Método	Var.Rem	Var.Omit	ROC	Sens	Spec	Dev	AIC
NVZ	134	22	92	85	85	515	617
COL	96	1	85	83	81	452	668
CORR	118	8	79	85	84	476	632
LASSO	158	12	93	85	86	500	568
PLS	154	6	79	85	84	758	820

Cuadro 5.2: Resumen métodos de regularización modelo logístico - a partir de 203 variables no agrupadas

Var.Rem Corresponde a las variables removidas por el método, *Var.Omit* variables omitidas en la creación del modelo de regresión. *ROC, Sens, Spec* al promedio de los

resultados del remuestreo, mientras que *DEV, AIC* a los resultados de la devianza y del criterio de akaike respectivamente. Agregando los resultados del Cuadro 5.3 se puede obtener una comparativa de los diferentes métodos de selección respecto al uso de las bases por separado.

Método	Var.Rem	ROC	Sens	Spec	Dev	AIC
AMB (36 variables)	23	92	84	82	668	696
SNP (167 variables)	155	67	60	63	1326	1352

Cuadro 5.3: Para la determinación de los modelos de cada base se utilizó selección de variables mediante LASSO, así como la limpieza por varianza a cero (NVZ) y correlación para los genotipos

La Figura 5.3 muestra la variabilidad de cada modelo en orden ascendente respecto al promedio de los resultados del remuestreo y a los modelos AMB y SNP, los cuales contiene solo variables de las bases *Base 1* y *Base 2* respectivamente.

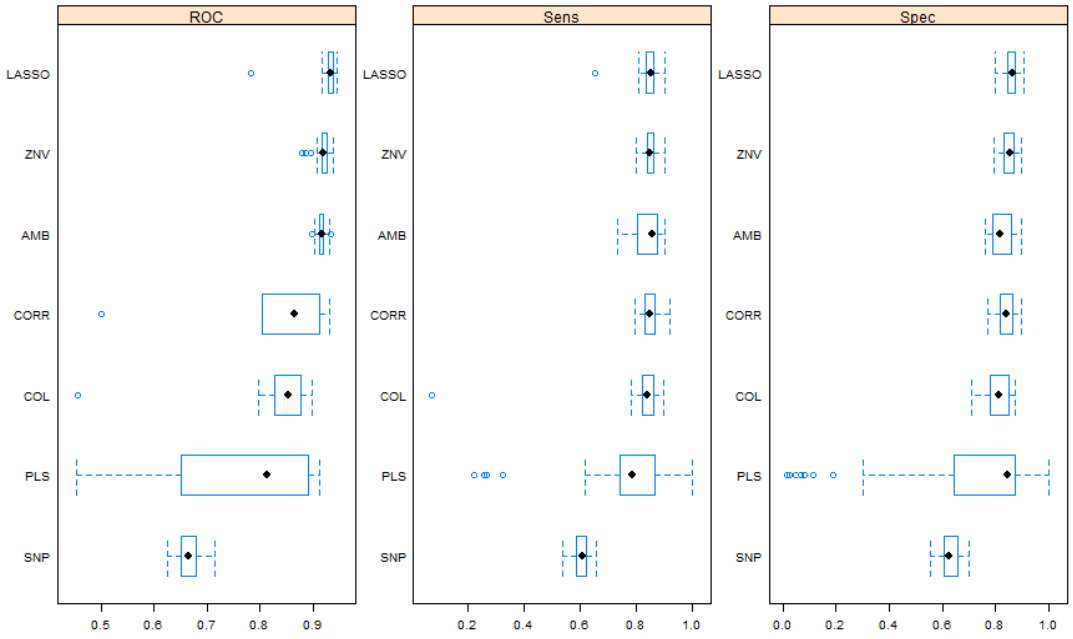


Figura 5.3: Comparativa métodos de regularización utilizados para el modelo logit

Particularmente en el caso de las variables propias de la *Base 1* los modelos **COL** y **COR** son las que retienen una mayor cantidad de estas, sin embargo como se puede

observar estos modelos también tienen mayor variabilidad en sus resultados, lo que motiva a considerar que las variables de la*Base 1* aumentan varianza al estar en conjunto con los genotipos. Por su parte, el modelo que solo incluye a 13 variables de la *Base 1* -modelo AMB- tiene una predicción alta respecto a los otros modelos, lo que podría indicar que son los genotipos los que introducen varianza.

	Logit - Factores genómicos			Logit - Factores ambientales		
	variable	estimador	p. valor	variable	estimador	p. valor
1	rs55885037.1TC	-1.47E+00	2.72E-05	Nage18_29	3.44E+00	7.46E-17
2	rs34151234.12GG	-1.11E+00	8.84E-05	Nage30_41	1.68E+00	7.96E-13
3	rs7205526.16CC	8.05E-01	6.09E-04	Nage55_68	-1.68E+00	5.65E-09
4	rs2598121.7CT	-1.15E+00	1.46E-03	FWCCSi	-2.63E+00	8.03E-09
5	rs2194310.16CC	-4.03E-01	3.72E-03	EdPrComp	-1.56E+00	8.53E-07
6	rs2194310.16TT	6.15E-01	8.68E-03	ObeSi	-1.45E+00	6.52E-06
7	rs74382455.1TT	8.68E-01	1.60E-02	Carne1.5	-1.18E+00	1.24E-04
8	(Intercept)	-1.46E+00	3.55E-02	SedNo	8.00E-01	2.00E-04
9	rs16868695.6CC	4.79E-01	5.77E-02	DiabSi	-1.84E+00	1.16E-03
10	rs7197593.16GG	5.74E-01	9.17E-02	(Intercept)	1.06E+00	1.34E-03
11	rs26934.3CC	4.02E-01	1.57E-01	EdEstpsg	1.66E+00	2.90E-03
12	rs2354487.2TT	-3.88E-01	2.49E-01	Alc75	-5.53E-01	1.22E-02
13	rs2354487.2CT	2.10E-01	5.46E-01	Alc75_150	6.73E-01	6.53E-02
14				PolSi	-1.83E+01	9.74E-01

Cuadro 5.4: Estimadores para los modelos logit para la Base 1 - Base 2.

Basado en el interés de observar las interrelaciones entre las diferentes variables se muestran los coeficientes y otras características del modelo determinado por método *LASSO* -Cuadro 5.5- El cual presenta el mejor desempeño.

El observar que para el modelo solo se muestren 3 genotipos del grupo de “Alta” contribución, 5 del de contribución “Media” y 9 de contribución “Baja” es indicador de la dificultad para modelar relaciones lineales entre los SNP’s. Los resultados anteriores pueden estar indicando que los SNP’s no son adecuadamente caracterizados por un modelo logístico a excepción de unos pocos asociados a los grupos de “Baja” contribución los cuales basado en la hipótesis asociación con la vía metiladora -Sección1.3.2- podrían justificar su efecto aditivo, así como sus asociaciones con las variables ambientales, hecho que se había anticipado en el Capítulo 4.

	variable	estimador	Std.Error	z.value	p.value	OR	sd.OR	LOR	UOR
1	Nage18_29	3.55E+00	5.27E-01	6.72E+00	1.77E-11	2.88E-02	5.90E-01	-1.13E+00	1.19E+00
2	Nage30_41	1.83E+00	2.93E-01	6.23E+00	4.55E-10	1.61E-01	7.46E-01	-1.30E+00	1.62E+00
3	FWCCNo	3.22E+00	5.32E-01	6.05E+00	1.49E-09	4.00E-02	5.87E-01	-1.11E+00	1.19E+00
4	Nage55_68	-2.03E+00	3.40E-01	-5.98E+00	2.26E-09	7.61E+00	7.12E-01	6.22E+00	9.01E+00
5	ObeNo	1.78E+00	3.90E-01	4.57E+00	4.81E-06	1.68E-01	6.77E-01	-1.16E+00	1.50E+00
6	EdPrComp	-1.67E+00	3.94E-01	-4.24E+00	2.23E-05	5.31E+00	6.74E-01	3.99E+00	6.64E+00
7	rs7205526.16CC	1.53E+00	4.09E-01	3.74E+00	1.84E-04	2.16E-01	6.64E-01	-1.09E+00	1.52E+00
8	SedSi	-9.70E-01	2.61E-01	-3.72E+00	2.01E-04	2.64E+00	7.70E-01	1.13E+00	4.15E+00
9	Carne1.5	-1.32E+00	3.60E-01	-3.66E+00	2.53E-04	3.73E+00	6.98E-01	2.36E+00	5.10E+00
10	Nage69_91	-3.79E+00	1.05E+00	-3.59E+00	3.26E-04	4.43E+01	3.48E-01	4.36E+01	4.49E+01
11	DiabNo	2.22E+00	6.73E-01	3.30E+00	9.57E-04	1.08E-01	5.10E-01	-8.92E-01	1.11E+00
12	rs74719289.4AG	2.17E+00	7.67E-01	2.83E+00	4.60E-03	1.14E-01	4.64E-01	-7.96E-01	1.02E+00
13	rs2194310.16CC	-6.89E-01	2.48E-01	-2.78E+00	5.42E-03	1.99E+00	7.81E-01	4.62E-01	3.52E+00
14	rs2598121.7CT	-1.45E+00	5.53E-01	-2.63E+00	8.63E-03	4.27E+00	5.75E-01	3.14E+00	5.40E+00
15	rs2194310.16TT	1.21E+00	4.84E-01	2.51E+00	1.22E-02	2.97E-01	6.16E-01	-9.11E-01	1.51E+00
16	Alc75_150	1.03E+00	4.77E-01	2.16E+00	3.09E-02	3.57E-01	6.20E-01	-8.59E-01	1.57E+00
17	rs2354487.2CT	1.25E+00	5.88E-01	2.12E+00	3.40E-02	2.88E-01	5.56E-01	-8.01E-01	1.38E+00
18	rs16868695.6CC	9.00E-01	4.27E-01	2.11E+00	3.51E-02	4.07E-01	6.52E-01	-8.72E-01	1.69E+00
19	EdEstpsg	1.28E+00	6.40E-01	2.00E+00	4.57E-02	2.79E-01	5.28E-01	-7.55E-01	1.31E+00
20	rs55885037.1CC	1.14E+00	6.39E-01	1.79E+00	7.41E-02	3.19E-01	5.28E-01	-7.15E-01	1.35E+00
21	rs7197593.16AG	-9.79E-01	6.31E-01	-1.55E+00	1.21E-01	2.66E+00	5.32E-01	1.62E+00	3.70E+00
22	rs115829688.1AG	-1.45E+00	9.79E-01	-1.48E+00	1.38E-01	4.28E+00	3.76E-01	3.54E+00	5.01E+00
23	Alc75	-3.58E-01	2.61E-01	-1.37E+00	1.70E-01	1.43E+00	7.70E-01	-7.93E-02	2.94E+00
24	rs118184226.12AG	-9.24E-01	8.00E-01	-1.15E+00	2.48E-01	2.52E+00	4.49E-01	1.64E+00	3.40E+00
25	rs26934.3CC	7.05E-01	6.10E-01	1.15E+00	2.48E-01	4.94E-01	5.43E-01	-5.70E-01	1.56E+00
26	rs57798805.14AC	9.67E-01	8.96E-01	1.08E+00	2.80E-01	3.80E-01	4.08E-01	-4.20E-01	1.18E+00
27	EdEUniv	3.00E-01	3.20E-01	9.36E-01	3.49E-01	7.41E-01	7.26E-01	-6.82E-01	2.16E+00
28	rs1797044.4CT	-5.87E-01	6.71E-01	-8.75E-01	3.81E-01	1.80E+00	5.11E-01	7.97E-01	2.80E+00
29	EdSecCom	-2.54E-01	3.09E-01	-8.21E-01	4.11E-01	1.29E+00	7.34E-01	-1.50E-01	2.73E+00
30	rs2354487.2TT	1.64E-01	5.57E-01	2.94E-01	7.68E-01	8.49E-01	5.73E-01	-2.75E-01	1.97E+00
31	PolNo	1.84E+01	5.23E+02	3.52E-02	9.72E-01	1.04E-08	1.14E-227	1.04E-08	1.04E-08
32	(Intercept)	-2.50E+01	6.52E+03	-3.83E-03	9.97E-01	7.00E+10	0.00E+00	7.00E+10	7.00E+10
33	rs61989760.14CT	-4.18E+00	6.54E+03	-6.39E-04	9.99E-01	6.54E+01	0.00E+00	6.54E+01	6.54E+01
34	rs61989760.14TT	-2.41E+00	6.54E+03	-3.68E-04	1.00E+00	1.11E+01	0.00E+00	1.11E+01	1.11E+01

Cuadro 5.5: Características modelo logit -LASSO sin agrupación de variables

Otro aspecto que llama la atención del modelo generado por *LASSO* es la posición de *EdPrComp* la cual corresponde a "Estudios de educación primaria completos". El hecho de que tener educación primara sea una desventaja en el padecimiento del CCR es discutible al no presentarse una asociación directa clara, sin embargo si consideramos la existencia de variables ocultas como el acceso a mejores recursos económicos o un mayor acceso y búsqueda de la información relacionada se puede justificar su presencia.

Cabe destacar que la ausencia de la variable *"No ed"* se ve justificada a partir de evitar la multicolinealidad. Es decir, es eliminada del modelo al ser presentada a partir del complemento del conjunto de las otras variables.

5.2. Otros métodos

Con el fin de modelar las relación complejas y hasta este punto no lineales determinadas por las asociaciones entre los SNP's y las variables de la *Base 1* se exploran otros métodos como Random forest(CFR), Adaboost (CAD), xtreme Gradient Boosting (CXGB), neural net (CNN) y soport vector machine (SVM), para cada uno de ellos se determinan los hiperparámetros a partir de validación cruzada.

5.2.1. Métodos de ensemble

Como se menciona en el Capítulo 2, estos métodos se basan en la combinación de estrategias, seleccionando y ponderando clasificadores débiles (por lo general árboles de decisión) con el fin de generar un clasificador robusto. Recordemos que hay 2 tipos de emsemble *Boosting* y *Bagging*.

CAD	CXGB	CRF
1. Profundidad del árbol (maxdepth) 2. Número de arboles (iter) 3. Factor de aprendizaje ν 4. Factor de importancia de cada árbol ω 5. entre otras muchas opciones dependiendo de la variante del algoritmo	1. Profundidad del árbol (max_depth) 2. Número de árboles (nrounds) 3. Factor de aprendizaje η 4. Error mínimo γ 5. Número mínimo de nodos (min_child_weigth) 6. Factor de remuestreo (subsample) 7. entre otras muchas opciones dependiendo de la variante del algoritmo	1. Número de nodos terminales (nodeSize) 2. Número de predictores por árbol (mty) 3. Número de árboles (ntree)

Cuadro 5.6: Hiperparámetros de los métodos de ensemble

Selección de hiperparámetros

En esta sección dentro de los métodos de boosting se consideran: Xtreme Gradient Boosting (CXGB) y Adaboost (CAD). La diferencia entre ambos métodos corresponde

al manejo de la función de penalización, el primero genera una función para cada árbol; mientras que Adaboost utiliza la misma para todos. Dentro de los métodos de bagging, utilizaremos Random Forest(CRF) que tiene un mejor manejo de la varianza evitando el sobre ajuste al ponderar cada árbol. La selección de los hiperparámetros se realiza mediante validación cruzada utilizando las librerías del paquete *Caret* al respecto el Cuadro 5.6 muestra algunas de las opciones a configurar.

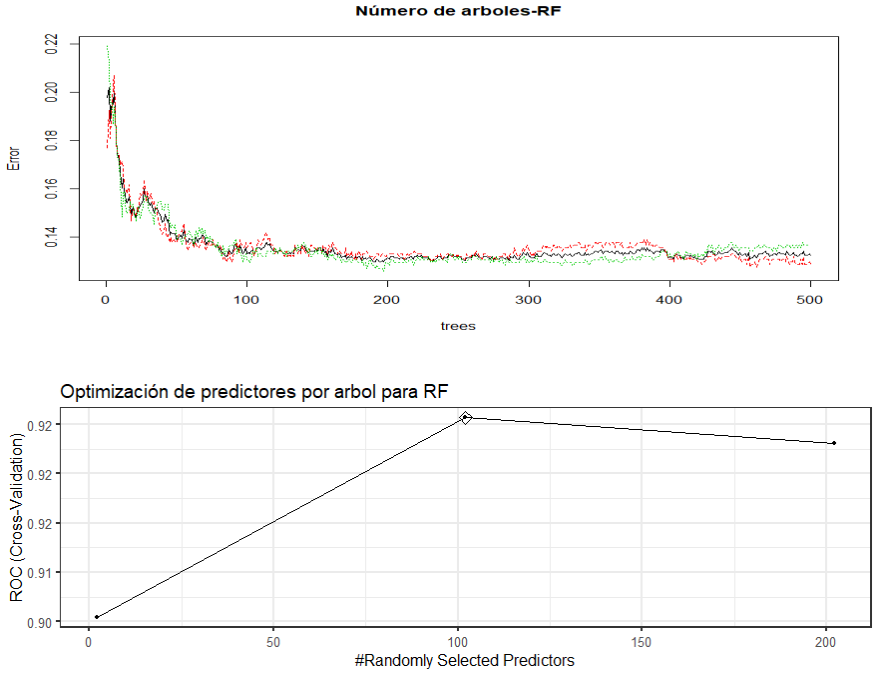
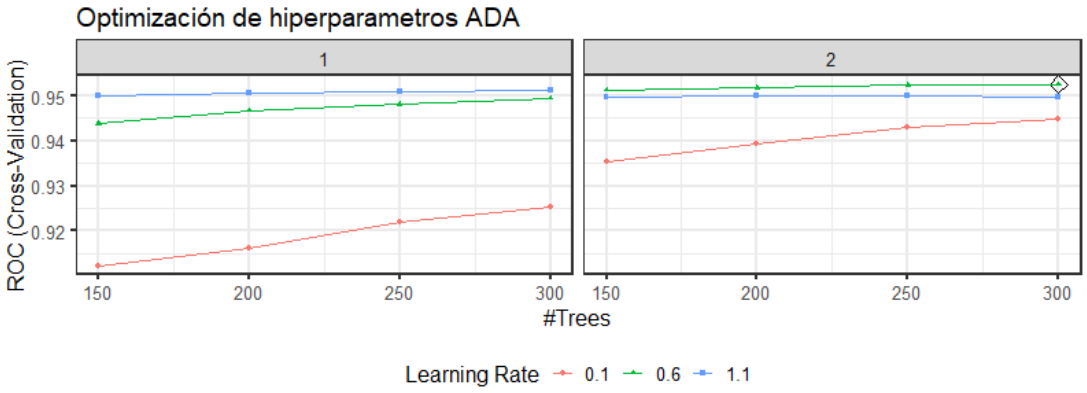
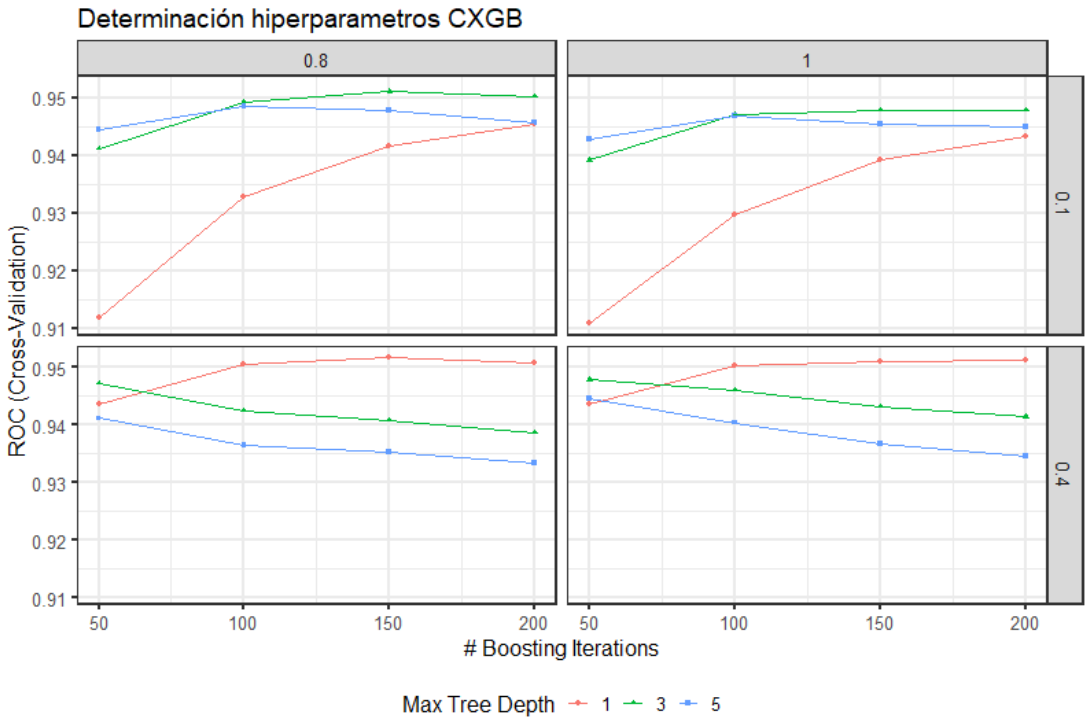


Figura 5.4: Como se puede observar a partir de 200 árboles el algoritmo se estabiliza, de forma predeterminada *caret* hace la búsqueda del número optimo desde 1 a 500 árboles, su re-definición no mejora los resultados solo disminuye el tiempo de cómputo, 102 predictores por árbol minimizan el error de clasificación.

Para los métodos de clasificación que involucran arboles de decisión se considera de forma predeterminada 1 como nodo terminal. En el caso de Random forest bajo el método *rf* de *caret*, el número de árboles (ntree) es 500 de forma predeterminada, este parámetro esta implicado en el coste computacional y se puede modificar a un valor menor sin repercutir esto en una mejora pero ahorrando recursos. Luego el único parámetro configurable es el número de predictores, el cual obtiene su óptimo en mtry= 102 como se observa en la Figura 5.4. Adicionalmente los diferentes colores representan el error de clasificación de cada clase, así como el error promedio.



(a) ***Adaboost***



(b) ***Xtreme gradient Boosting***

Figura 5.5: Hiperparámetros para métodos de Boosting

En el caso de Adaboost mediante el método *ada*, se tienen como parámetros a optimizar la profundidad, el número de árboles y el factor de aprendizaje *Caret* de forma predeterminada realiza una búsqueda del óptimo con los parámetros $\nu = \{0{,}1\}$, $iter = \{50, 100, 150\}$, $maxdepth = \{1, 2, 3\}$ los cuales sirven de base, a partir de estos se

actualiza la búsqueda a $\nu = \{\frac{1}{2}, 1, 2\}$,$maxdepth$= 1, $iter$=$\{100, 175, 250\}$, la cual genera como resultados los óptimos al tomar ν=0.6 $maxdepth$ = 2 y $iter$ = 300.

Para el caso de CXGB se optimizan los parámetros sobre el número de árboles, profundidad de cada árbol, factor de aprendizaje y porcentaje de submuestreo obteniendo como valores óptimos $nrounds$ = 150, η = 0.4 y $subsample$ = 0.8. Los entrenamientos para los métodos de boosting se muestran en la Figura 5.5.

5.2.2. Redes Neuronales (NN)

La redes Neuronales son muy utilizados en la industria producto de sus buenos resultados, los cuales dependen de tamaño del conjunto de datos de entrenamiento para que la red "aprenda" bien. En muchos casos es difícil interpretar las ponderaciones generadas por la red. El método *nnet* de *Caret* tiene dos parámetros a entrenar *size* que refiere al número de capas ocultas y *decay* el cual corresponde al factor de aprendizaje. De forma predeterminada realiza 100 iteraciones por capa hasta alcanzar convergencia, la Figura 5.6 muestra la determinación de los valores óptimos *size*=1, *decay*=0.01.

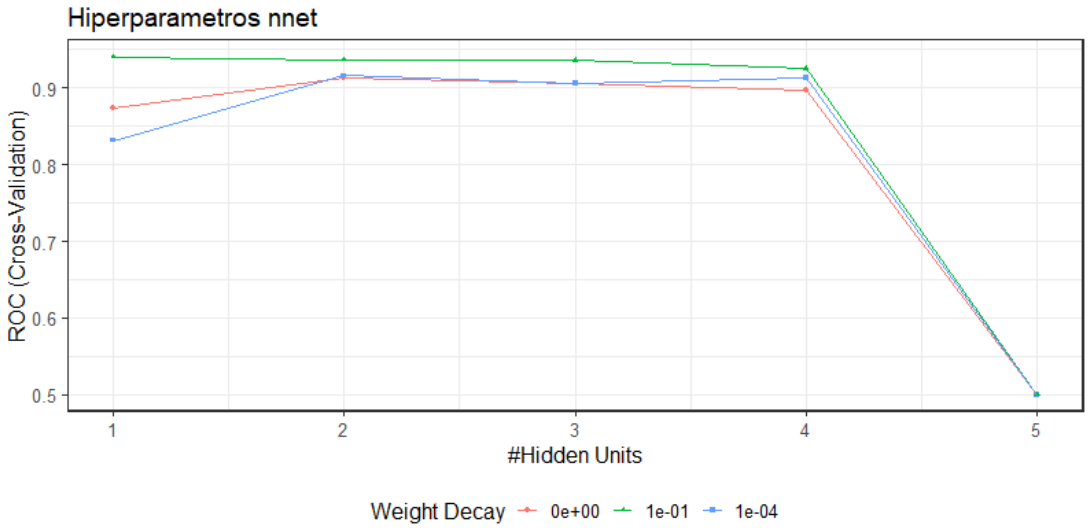


Figura 5.6: Hiperparámetros CNN

5.2.3. Maquinas de Vectores de Soporte (SVM)

Finalmente, basado en la idea de que los datos son linealmente separables bajo el supuesto de relación lineal - aditiva, se propone el uso de un plano separador mediante el uso de SVM. El método *svmLinearWeights* de *Caret* incorpora 2 parámetros *costo* y *peso*, los cuales luego de realizar la búsqueda corresponden a *c=0.5* y peso *weight = 1*. Por otro lado, dado que en los modelos de regresión logit se observo que los datos

no se adecuaban “”bien” a un modelo lineal, se ajustan también los parámetros para un modelo con kernel radial, el cual es generado mediante el método *svmRadial*. para este los parámetros de ajuste son *C=0.02* de costo y el parámetro *sigma=1.2* el cual corresponde a la penalización. El entrenamiento se muestra en la Figura 5.7.

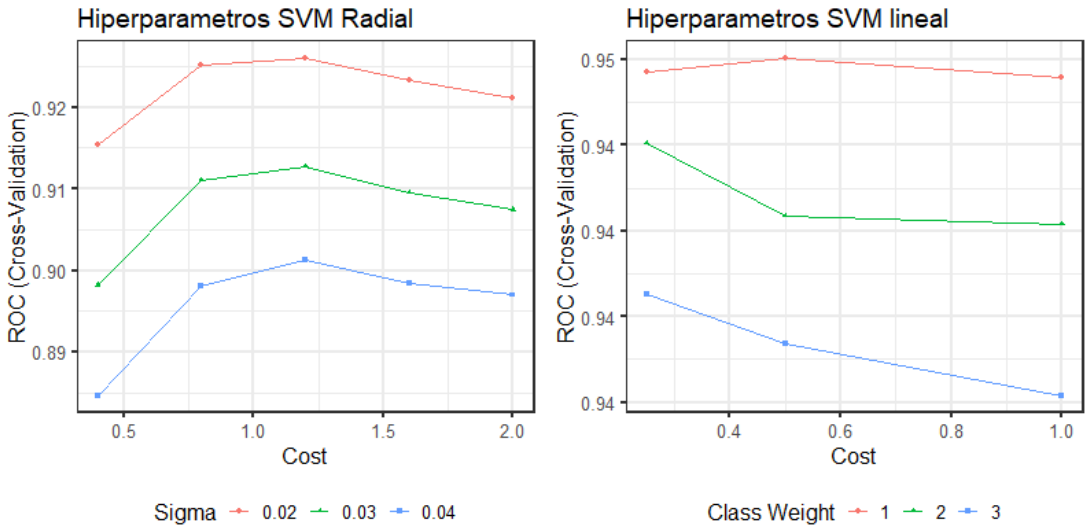


Figura 5.7: Hiperparámetros SVM

5.2.4. Comparativa

Para los métodos considerados anteriormente, el Cuadro 5.7 resume los resultados. En particular el modelo generado por XGB con un ROC $\approx 95\,\%$ tiene características de interés que se discutirán más adelante.

	ROC.Median	ROC.Mean	Sens.Median	Sens.Mean	Spec.Median	Spec.Mean
RF	0.92	0.93	0.86	0.86	0.86	0.86
ADA	0.95	0.95	0.88	0.88	0.89	0.89
XGB	0.96	0.95	0.88	0.88	0.89	0.89
NN	0.94	0.94	0.86	0.86	0.88	0.88
SVMlineal	0.95	0.95	0.88	0.87	0.90	0.89
SVMRadial	0.92	0.93	0.88	0.88	0.84	0.84

Cuadro 5.7: Comparativa calidad de los estimadores M.L.

Como se puede observar, la calidad de la predicción de los métodos de aprendizaje automático en el peor de los casos es igual al obtenido por el modelo logístico usando

LASSO, dando idea clara de una mejor interpretación de la asociación entre las variables.

Una característica a analizar en cada modelo, es el manejo que realiza el algoritmo sobre las categorías de una misma variable; al respecto se observa que RF, AD, SVM tienden a priorizar las variables con más niveles, XGB por su parte, parece reconocer las características de una misma categoría realizando un balance entre la varianza generada al eliminarla y el poder predictivo lo cual es resultado de su búsqueda de ajuste; esto es conveniente al no poder realizar un pre-procesamiento por el peligro de eliminar posibles relaciones de interés.

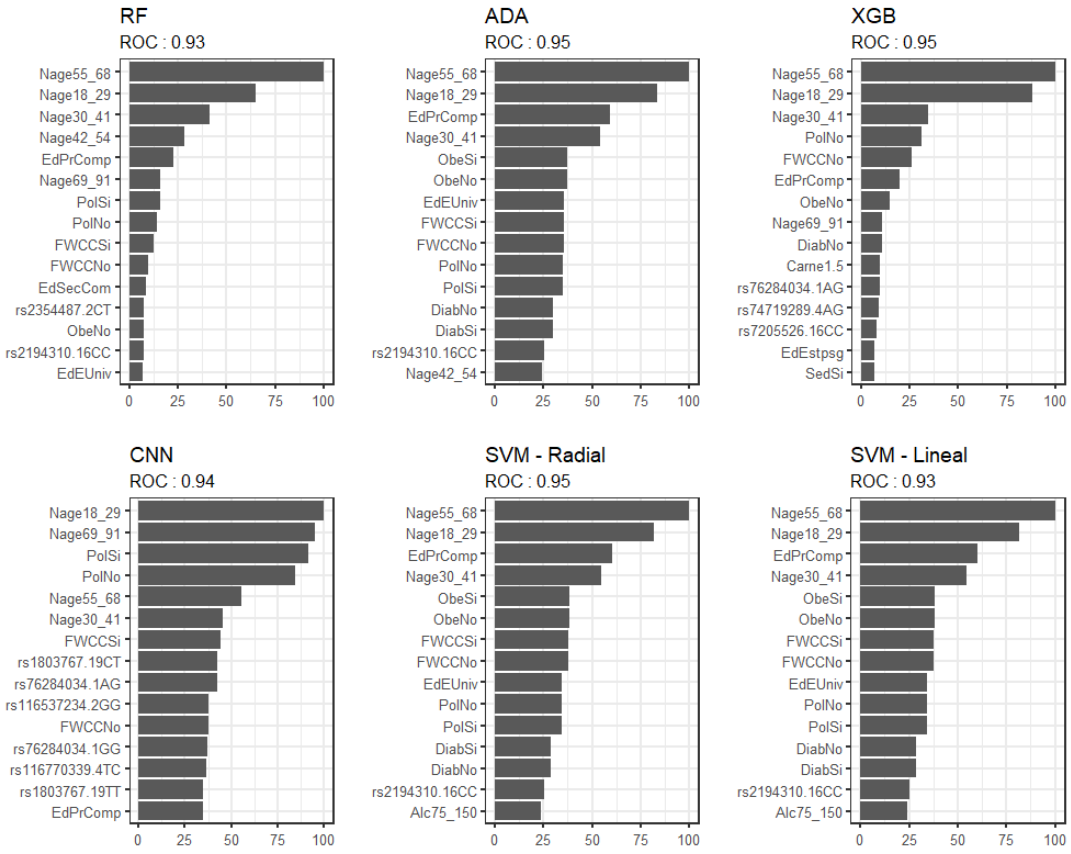


Figura 5.8: Se muestran las 15 variable principales de las 203 determinadas por los métodos de Aprendizaje máquina.

Estas consideraciones surgen de la observaciones de las 15 variables principales de cada método; las cuales se presentan en la Figura 5.8; En la figura los predictores son ordenados en forma descendente de acuerdo a su contribución a la curva ROC, en el eje X se ha realizado una ponderación para comparar la importancia relativa de cada variable, respecto a la variable más informativa.

Claramente, los diferentes métodos dan importancias distintas a una misma variable; sin embargo en general se inclinan a las variables de la *Base 1* -particularmente a *Nage*- reduciendo la participación de los genotipos; llama la atención el comportamiento ligeramente atípico que muestran las redes neuronales (CNN) al incluir un mayor número de genotipos y dar una importancia menor a *EdPrComp*.

Previamente en la sección 3.1.2 las variables *Nage* y *Ed* mostraron una fuerte asociación con el diagnostico del CCR, esto a partir del comportamiento monótono de algunas de sus categorías -Figura5.9-.

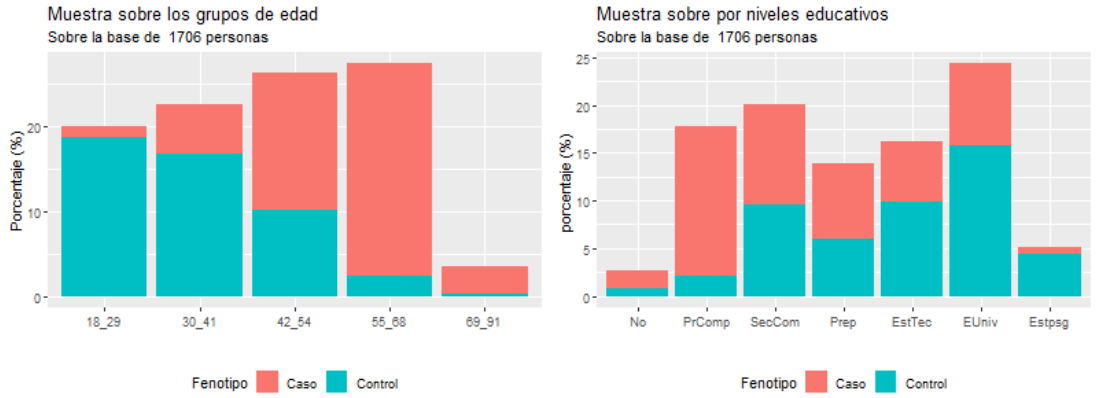


Figura 5.9: Detalle de frecuencias relativas para variables destacadas

El comportamiento a la alza en la cantidad de casos conforme aumenta la edad es un comportamiento típico presente en el CCR, lo que hace inclinar la atención a una segregación de las variables, particularmente por grupo de edad. Las variables pólipos *Pol* y historial familiar *FWCC* son eliminadas bajo el supuesto de que sus relaciones son mostradas mediante los SNP's.

5.3. Estratificación por grupos de edad

Se consideran los cuatro primeros grupos los cual muestran una mayor homgeneidad en el número de datos. Se entrenan los algoritmos de XGB sobre ADA y RF -por su velocidad de calculo- CNN -por su inclusión de genotipos- así como el modelo logístico con *LASSO* el cual mantiene las relaciones interés entre los SNP.

Para este caso, la selección de los hiperparametros se hace sobre la opción predeterminada de *caret* con la función*TuneLength* = 3, con lo que se hace una selección aleatoria sobre tres valores para cada posible parámetro a configurar. De forma empírica se ha evidenciado que este método da buenos resultados.

Las variables de importancia generadas por cada grupo de edades se escalan y posteriormente son promediadas, los resultados de los 15 primeros indicadores de cada grupo se muestran en el Cuadro 5.8

	Orden 18-29				
	variable	X18_29	X30_41	X42_54	X55_68
1	Smok5	**5.83**	0.40	0.44	8.42
2	SmokNo	**5.59**	3.58	2.37	0.09
3	rs34151234.12AG	**5.58**	0.27	0.19	0.14
4	SedSi	**5.49**	0.91	3.40	4.65
5	EdSecCom	**5.20**	1.39	0.46	1.17
6	GenerM	**4.36**	10.29	0.55	0.02
7	rs11860295.16CC	**4.22**	0.44	1.93	1.48
8	rs16868695.6CC	**3.73**	0.39	2.69	0.01
9	rs2194310.16CC	**3.50**	0.89	4.01	0.14
10	rs116847323.3CC	**3.37**	0.53	0.16	0.04
11	rs2354487.2CT	**3.16**	0.65	4.32	2.57
12	rs8052655.16AG	**3.12**	0.13	0.03	0.12
13	ObeNo	**3.02**	4.86	6.04	0.18
14	EdPrep	**2.07**	0.82	0.75	8.74
15	Alc75_150	**2.06**	2.02	4.30	0.31

	Orden 30-41				
	variable	X18_29	X30_41	X42_54	X55_68
1	rs116017843.2CC	0.11	**17.59**	0.39	0.71
2	GenerM	4.36	**10.29**	0.55	0.02
3	ObeNo	3.02	**4.86**	6.04	0.18
4	ObeSi	0.10	**4.21**	0.65	0.28
5	SmokNo	5.59	**3.58**	2.37	0.09
6	EdPrComp	0.02	**3.42**	6.16	9.08
7	rs76528184.18CC	0.23	**2.22**	0.02	0.07
8	rs76284034.1AG	0.02	**2.20**	0.07	0.37
9	Alc75_150	2.06	**2.02**	4.30	0.31
10	rs76528184.18TC	0.13	**1.79**	0.06	0.03
11	EdEstpsg	0.36	**1.59**	2.24	1.52
12	Alc75	0.66	**1.55**	0.62	4.97
13	Carne1.5	1.06	**1.48**	4.03	6.48
14	rs192567137.5CC	0.13	**1.47**	0.04	0.09
15	EdSecCom	5.20	**1.39**	0.46	1.17

	Orden 42-54				
	variable	X18_29	X30_41	X42_54	X55_68
1	EdPrComp	0.02	3.42	**6.16**	9.08
2	ObeNo	3.02	4.86	**6.04**	0.18
3	DiabNo	0.12	0.05	**5.08**	3.42
4	rs2354487.2CT	3.16	0.65	**4.32**	2.57
5	Alc75_150	2.06	2.02	**4.30**	0.31
6	rs2598121.7CT	0.03	0.10	**4.12**	0.14
7	rs115490452.6CC	0.13	0.06	**4.11**	0.01
8	Carne1.5	1.06	1.48	**4.03**	6.48
9	rs2194310.16CC	3.50	0.89	**4.01**	0.14
10	SedSi	5.49	0.91	**3.40**	4.65
11	rs16868695.6CC	3.73	0.39	**2.69**	0.01
12	SmokNo	5.59	3.58	**2.37**	0.09
13	EdEstpsg	0.36	1.59	**2.24**	1.52
14	rs11860295.16CC	4.22	0.44	**1.93**	1.48
15	rs55885037.1CC	0.11	0.34	**1.87**	0.06

	Order 55-68				
	variable	X18_29	X30_41	X42_54	X55_68
1	EdPrComp	0.02	3.42	6.16	**9.08**
2	EdPrep	2.07	0.82	0.75	**8.74**
3	Smok5	5.83	0.40	0.44	**8.42**
4	Carne1.5	1.06	1.48	4.03	**6.48**
5	Alc75	0.66	1.55	0.62	**4.97**
6	CarneNo	0.19	0.21	1.03	**4.90**
7	SedSi	5.49	0.91	3.40	**4.65**
8	rs2354487.2TT	0.30	1.21	0.94	**3.71**
9	DiabNo	0.12	0.05	5.08	**3.42**
10	DiabSi	0.06	0.05	0.58	**3.27**
11	rs8047080.16AA	0.37	0.11	0.15	**3.20**
12	rs74382455.1CT	0.57	0.58	0.50	**3.11**
13	rs2354487.2CT	3.16	0.65	4.32	**2.57**
14	rs1797044.4CT	0.04	0.09	0.58	**1.79**
15	EdEstpsg	0.36	1.59	2.24	**1.52**

Cuadro 5.8: El orden en cada cuadro muestra la importancia de las diferentes variables de acuerdo a los diferentes grupos de edad.

La segmentación para los grupos de edad, muestra que para cada variable se da una ventana de importancia, la cual a lo largo del tiempo va cambiando, en particular se observa que la importancia que van tomando las variables ambientales aumenta a

lo largo de la edad, al tiempo que disminuye la importancia relativa de los genotipos. Si bien la contribución global de los SNP se mantiene estable -alrededor de un 55 %, excepto para el grupo 55-68 para el cual baja a un 40 %-, la muestra tomada en los cuadros, indica que participación disminuye pasando de un 25 % en el grupo de edad 18-29 a un 14.38 % en el grupo de 55-68 lo que conlleva a pensar que hay un mayor número de SNP's participantes pero de menor contribución.

Esto justificaría en parte la dificultad para incorporar a los SNP's en los diferentes modelos que incluyen a los rangos de edad. Finalmente el Cuadro 5.9 muestra los resultados de evaluación de los diferentes modelos ante valores nuevos. Para cada modelo se utiliza validación cruzada con remuestreo utilizando los datos generados de filtrar la *Base Final* por cada uno de los grupos de edad.

	ROC - real	Linf	Lsup	ROC-entrenamiento
CGLM18_29	0.30	0.25	0.35	49
CXGB18_29	0.18	0.14	0.23	80
CNN18_29	0.18	0.14	0.22	75
CGLM30_41	0.36	0.31	0.42	66
CXGB30_41	0.45	0.40	0.51	81
CNN30_41	0.47	0.41	0.53	73
CGLM42_54	0.83	0.78	0.87	79
CXGB42_54	0.82	0.78	0.86	81
CNN42_54	0.79	0.74	0.84	79
CGLM55_68	0.89	0.84	0.92	82
CXGB55_68	0.90	0.86	0.93	81
CNN55_68	0.88	0.84	0.92	79

Cuadro 5.9: El cuadro muestra los resultados comparativos de los diferentes métodos ante un conjunto de prueba independiente.

La potencia de los métodos de aprendizaje automático es clara, obteniendo buenos resultados a pesar de tener conjuntos de entrenamiento limitados, sin embargo al considerar el costo computacional que conlleva, adicional a la pérdida de interpretación hacen pensar que su uso debe ser prudente.

El Cuadro 5.10 muestra una serie de genotipos y sus características que se han recopilado a lo largo del trabajo. Estos presentan relaciones de interés al corresponder a variables asociadas a genes o oncogenes. Las características mostradas son tomadas del Centro Nacional de Información Biotecnología NCBI.

Genotipo	Características génicas	Asociaciones
rs2194310.16CC	asociado al gen LINC02141	Esta presente en los modelo 13-Lasso, 14 -Rf,14-Ada, 14 - SvmRadial, 14-SvmLineal. los genotipo TT, CC se presentaban en el grupo de contribución *N.Bajo* mostrando interrelaciones con variables ambientales no cumple equilibrio HW.
rs115829688.1AG	asociado dos oncogenes MINOS1-NBL1 y NBL1	El genotipo AG aumenta susceptibilidad se presenta en 22-lasso
rs116770339.4TC	codificador STK32B mutaciones están asociadas a enfermedades oseas	propuesto como protector en genotipo TC y generador de susceptibilidad en CC. 13-nnet
rs7205526.16CC	gen codificador LRRC36 Leucina	13 -xgb, 7-lasso, el genotipo CC se presentaba en el grupo de contribución *N.Bajo* mostrando interrelaciones con variables ambientales.
rs74719289.4AG	oncogen SCARB2	el genotipo AG es protector ,12-xgb, 12-lasso, N.medio, presente en HW con un OR sobre saliente
rs2354487.2CT		presente en 12-RF y 17-Lasso su importancia aumento al realizarse agrupación de variables, se presento en los grupos contributivos *N.Bajo* en el cual mostraba asociaciones con variables ambientales. No cumplia el equilibrio de HW
rs116537234.2GG		Presente en HW con un OR sobre saliente, el genotipo AG se propone como protector GG como precursor 10-nnet.
rs16868695.6CC	genes codificadores SCUBE3 peptido - oseo LOC101929285 desconocido	18- lasso
rs55885037.1CC	gen LINC01759" regulador	el genotipo TC se ubica en el grupo de contribución media 20-lasso
rs26934.3CC	gen codificador PRICKL2 funcion desconocida asociado al sistema nervioso y a retinoblastoma	25 -lasso
rs61989760.14TT	gen RPS6KA5" codificador de kinasa encargado de la regulación del fosfato	esta en TC en el grupo Alto 24-lasso
rs1803767.19	Asociado al oncogen Gene: DNAJB1, estimula la actividad de las células	8-nnet
rs76284034.1AG	asociado al gen codificador ACBD6 regulación de lipidos.	presente en 11-xgb, 9-nnet
rs76284034.1AG	codificador gen ACDB6	11-xgb, 9-nnet

Cuadro 5.10: Características de los genotipos frecuentes.

Capítulo 6

Conclusiones y futuros trabajos

En el presente trabajo los Capítulos 1 y 2 son introductorios y exponen generalidades sobre los conceptos a desarrollar. En el Capítulo 1 se hace un breve resumen sobre trabajos previos entorno al cáncer en general y al CCR en especifico, introduciendo conceptos técnicos propios de la genómica; por su parte el Capítulo 2 presenta referencias sobre los métodos de ciencia de datos aplicados, dando una breve descripción de cada uno.

El Capítulo 3 describe algunas de las dificultades referentes al proyecto, la limpieza de la muestra y la combinación de las bases, pasando de tener aproximadamente 3525 observaciones a 1707. En la sección 3.1.2 se realiza una caracterización de las edades con el fin de trabajar todas la variables como categóricas manteniendo el valor predictivo.

Se generan asociaciones entre las variables no genotípicas mediante un escalamiento multidimensional -Figura 3.7- el cual muestra disimilitudes en las proyecciones para los casos y controles particularmente para variables reconocidas históricamente como protectoras.

Se muestra vía correlación tetracórica -Figura 3.5- interacciones entre variables de la *Base 1* con el fenotipo-caso como las principales; así como relaciones entre los niveles de diferentes variables *EdNo* $\sim$ *CarneNo*, *EdEstPosg* $\sim$ *Smok20*, *Alc75-150*, *SmokNo*$\sim$*AlcNo*, *Diab*$\sim$*Nage+55* mostrando evidencia de una población con clases marcadas por el genero.

Por otro lado, el análisis de la *Base 2* permite determinar grupos de genotipos asociados a los *Casos* y *Controles* mediante el mapa de calor presente en la Figura 3.8, los cuales posteriormente se asocian con los genotipos determinados como protectores o precursores.

Se culmina el capítulo proponiendo modelo híbrido generado a partir de considerar binarizar los genotipos y aplicar selección de variables, esto en razón de conservar la

dependencias entre SNP's, las cuales generan diferentes efectos respecto a su análisis individual.

Los capítulos anteriores completan el análisis exploratorio, el Capítulo 4 abarca los objetivos, e inicia con la introducción algunas ventajas del análisis MCA respecto a PCA; entre ellas:

- Mejor representatividad de los datos en las primeras 2 dimensiones,
- Ganancia en la descripción de la inercia -equivalente a varianza- a lo largo de las 10 dimensiones consideradas.
- Interpretación directa de variables asociadas o no asociadas.

Se encuentra vía análisis de correspondencia múltiple (MCA) la existencia de variables asociadas entre sí y su efecto conjunto, al tiempo que se eliminan aquellas cuya presencia genera un efecto confusor en el análisis reduciendo la inercia -Cuadro 4.1 y 4.2-.

De los genotipos eliminados mediante el análisis de contribución se observa que 10 de los 12 corresponden a los SNP's poco significativos del gráfico Manhattan Figura 3.9 representando un medio adicional para reconocer SNP's no relevantes.

Se observa que las variables de la *Base 1*; tiene un fuerte efecto predictivo al tiempo que se distribuye a lo largo de las diferentes dimensiones y que se presentan principalmente partir de la 5ta, lo que equivale a decir que estas no tienen un efecto significativo de forma individual, sin embargo en conjunto permiten describir un aproximado de 35 % de la información -ver Figura 4.4-; respecto a los genotipos se interpreta que el bajo nivel predictivo es debido a que una asociación lineal no permite identificar el comportamiento complejo de los SNP's. hecho que se pone en evidencia en el Capitulo 5 al observar los resultados de los métodos de aprendizaje de maquina.

El comportamiento de los *genotipos* vs las variables no *genotípicas* -Figura 4.3- hacen ver que el efecto de las segundas es gradual y escaso en un pequeño lapso, el cual determina que el CCR tiene en un principio un origen genómico el cual posteriormente da lugar a las repercusiones por causas histórico - familiares, culminando en causas no genómicas Figura 4.8. Al respecto, se establecen grupos contributivos -Cuadro 4.5- los cuales muestran asociaciones con las diferentes etapas del desarrollo del CCR.

Los grupos contributivos corresponden a una ponderación de la contribución global de cada una de las variables a la explicación de la inercia -equivalente a la varianza en MCA-.

Puntualizando, los objetivos obtenidos en este capitulo corresponden a:

- Establecer las asociaciones entre las variables genómicas y ambientales con el CCR, indicando los porcentajes de contribución por parte de cada una 65 % - 35 % respectivamente.

- Explicitar las asociaciones entre las variables a partir de las distancias χ^2 mostrando los diferentes niveles de importancia de los SNP's a partir de la creación de grupos contributivos.
- Mostrar que efectivamente los porcentajes de participación para las variables genómicas y ambientales coinciden con los conocidos.

Basado en el interés de que suscito el estudio, se agrega un último capitulo de modelos predictivos el cual complementa los resultados del Capitulo 4. En este se considera realizar una agrupación de variables, a partir de los resultados de las Figuras 3.7 y 4.1.

- Est_Basico$\sim$ EdPrComp + EdNo
- Est_Medio $\sim$ EdEstpgs + EdEstTec + EdEUniv
- Est_Esp $\sim$ EdSecCom + EdPrep
- SmokSi $\sim$ Smok5 + Smok5_20 + Smok20
- AlcSi $\sim$ Alc75 + Alc75_150 + Alc150
- CarneSi $\sim$ Carne1.5 + Carne1.5_4.5 + Carne4.5

Sin embargo en el Capítulo 5 se descarta tal agrupación al reducir la significancia, introducir variabilidad y generar falta de ajuste restando poder predictivo en los modelos logísticos creados. Las curvas de aprendizaje -Figura 5.1- y la comparativa de los modelos son muestra de esto.

LASSO - ContrB		*LASSO - ContrM*	*LASSO - ContrBA*
7- *rs7205526.16CC*	20- *rs55885037.1CC*	12- *rs74719289.4AG*	22- *rs115829688.1AG*
13- *rs2194310.16CC*	25- *rs26934.3CC*	14- *rs2598121.7CT*	26- *rs57798805.14AC*
15- *rs2194310.16TT*		21- *rs7197593.16AG*	33- *rs61989760.14CT*
17- *rs2354487.2CT*	30- *rs2354487.2TT*	24- *rs118184226.12AG*	
18- *rs16868695.6CC*	34- *rs61989760.14TT*	28- *rs1797044.4CT*	

Cuadro 6.1: Intersección del Modelo *LASSO* con los grupos contributivos.

El Cuadro 6.1 muestra SNP's de interés al corresponder a la intersección de los resultados del análisis MCA en conjunto con los modelos logísticos identificando variables que en ambos modelos han resultado significativas, muestra particularmente genotipos

asociados a genes o oncogenes los cuales corresponden con un efecto directo hacia alguna patología. No obstante, estas asociaciones no tienen repercusiones actualmente estudiadas con el CCR, esto según la información consultada en el Centro Nacional de Información Biotecnología NCBI.

Por otro lado, otro de los resultados que permiten generar los modelos logísticos, son el análisis de los odd ratio, o razón de la ventaja al comparar individuos portadores de una característica, respecto a los no portadores. Considerando que en los modelos se considero como variable de interés la tenencia de CCR, se genera la Figura 6.1.

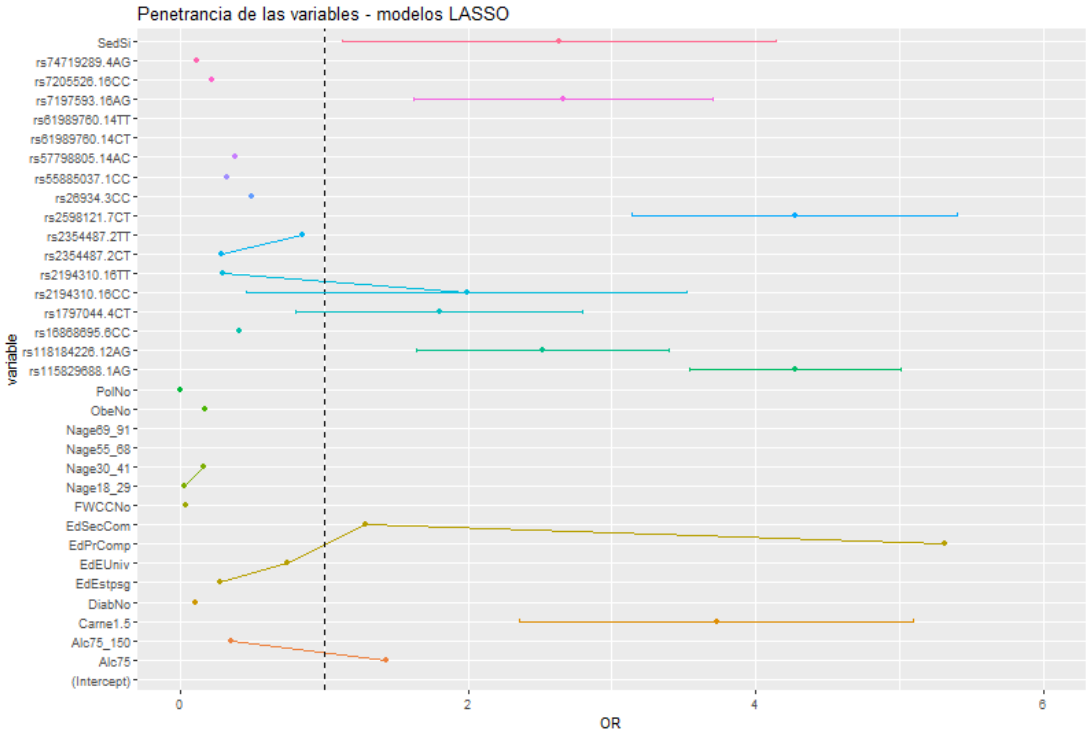


Figura 6.1: Odd-Ratio modelo logístico.

En ella se representan las variables del modelo logístico indicado en el Cuadro 5.5. La linea punteada corresponde a $OR = 1$ indicador de una asociación no determinante. Se consideran $OR < 1$ como variables asociadas a un efecto protector, mientras que $OR > 1$ como variables asociadas a un efecto causal. El punto corresponde con el valor de la OR, mientra que los margenes a los intervalos de confianza al 95 %. En algunos se muestran lineas asociadas a varias variables representando niveles de una misma variable.

Son particularmente interesante variables como *"rs2599121.7CT"*,*"rs11582688.1AG"* al corresponder a $OR > 4$ lo cual de acuerdo al criterio profesional del asesor del área

es *inusualmente alto*. Caso similar con los niveles educativos en cuyo caso se considera esta existiendo un efecto producto de la interacción entre niveles.

Respecto a los métodos de aprendizaje automático como Random Forest (RF), adaboost (Ada), XtremeGradient boosting (XGB) muestran una preferencia hacia las variables de la *Base 1* al estar esta presentes en una mayoría de las divisiones de los nodos, la Figura 6.2 muestra esta preferencia, ubicando pocos o muy pocos SNP entre las primeras posiciones de importancia para cada modelo.

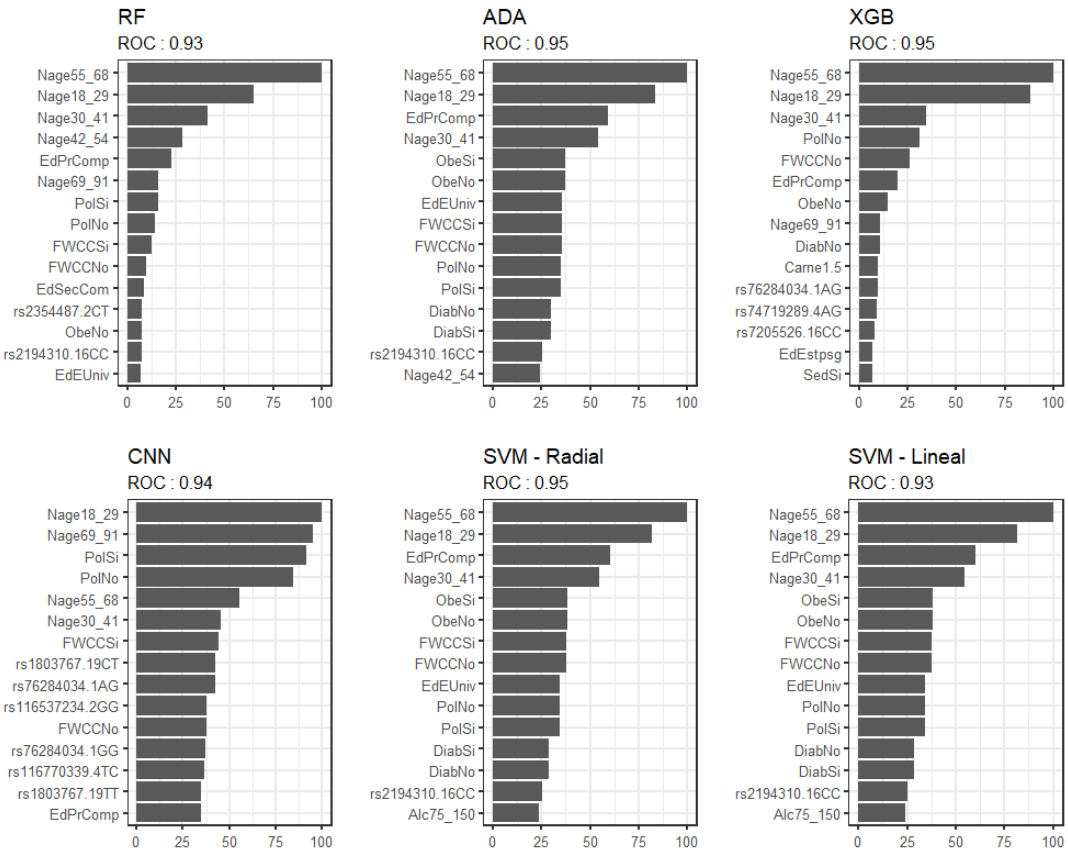


Figura 6.2: Comparativa de variables contributivas por modelo.

Para los diferentes conjuntos se observan los porcentajes descriptivos por base - Cuadro 6.2 - donde Adaboost (Ada), Neural Net(NN), Soport Vector Machine(SVM) son los métodos que se aproximan a los porcentajes de participación tradicionalmente establecidos (70 genético - 30 ambiental)

	Rf	Ada	NN	SVM	XGB	GLM
Base 1	76.5	29.9	31.5	30.7	84.5	66.3
Base 2	23.4	61.1	68.5	69.3	15.5	33.7

Cuadro 6.2: Porcentaje descriptivo por método y base.

De las Figuras 6.2 y 6.2 se concluye que en dependencia del objetivo de trabajo, RF,XGB,GLM son los algoritmos más adecuados por precisión, así como representatividad de los datos; sin embargo SVM, Ada, SVM son los que tienen resultados más cercanos a los presentes en la literatura.

Al final del Capítulo 5 se crearon estratos en los niveles de edad y se eliminaron variables de tipo genético -Pólipos, historial familiar- con el fin de observar las relaciones de las variables ante cada grupo, estas asociaciones determinaron que la importancia del efecto de una variable cambia con el tiempo reduciéndose en el caso de la mayoría de los SNP y aumentan en el caso de las variables no genéticas; este hecho amplia el panorama de estudio al indicar que quizá sea más adecuado el análisis de un mismo rango de edad, el cual podría dar resultados más asertivos, respecto al estudio de una población en general.

6.1. Futuros trabajos

- De la fase de exploración surgen algunas preguntas, en particular se plantea la hipótesis: Está la Leptina asociada con la génesis del CCR?. Esto por ser una proteína asociada al metabolismo de las grasas, mismas que corresponden con una fuente de energía. El aumento descontrolado de el número de células. Se plantea que el crecimiento celular requiere de un mayor componente energético y por tanto este se podria ver reflejado en mayores niveles de Leptina. En general para todos los tipos de cáncer y en particular para el CCR como se lee en la sección 1.3.1. La pregunta a partir del análisis de varios SNP generadores de aumento de la susceptibilidad al CCR determinado.

- Las asociaciones entre los grupos de variables del Capítulo 4 con las vías desarrollo del CCR, las cuales muestran similitudes a partir de las características de la penetrancia que las conforman, así como los porcentajes de participación. El explorar la asociación de los SNP con las diferentes vías, podría revelar asociaciones no observadas en SNP de baja penetrancia.

- La variable nivel educativo, muestra resultados de interés, al corresponder con un factor discriminador entre los casos y controles. Sin embargo, los datos no per-

miten obtener conclusiones sobre el motivo de estos resultados. Esto al presentar una aparente independencia con las otras variables en estudio. Del mismo modo la escasa participación de los niveles de consumo de carne son atípicos a los resultados conocidos, lo que podría motivar un estudio mas profundo al respecto.

- El análisis realizado es de tipo genómico, es de interés observar si estos resultados se replican en un análisis alélico.

- Los porcentajes de participación obtenidos en el Cuadro 6.2 no concuerdan con lo indicado en la literatura [33], de hecho los modelos que mejor predicen inclinan la balanza mayoritariamente a variables no genómicas, determinar si los resultados son producto de errores de procedimiento, de muestreo, entre otros o si son resultado de características particulares de la población en estudio que la hace más susceptible al CCR a partir de la exposición a los factores no génicos es un punto a investigar.

- Los alcances del presente trabajo limitan la aplicación de estrategias que podrían resultar de interés; como el **análisis factorial**. Mediante este, se podrían sugerir grupos de SNP's y valorar su aporte en conjunto. Actualmente los resultados considerados no toman en cuenta estas aportaciones de forma explicita limitando el alcance para ciertos grupos de baja penetrancia.

Finalmente, lejos de ser un estudio acabado, se requiere de la valides de expertos en el área; esto particularmente por la falta de información respecto a las asociaciones observadas; lo que demuestra que es primordial contar con equipos de trabajo multidiplinarios que nutran la investigación. Esto por cuanto se requieren conocimientos de varias áreas, entre ellas: biología, matemática, computación, genética, estadística, entre otras.

Glosario

Adenocarcinoma

Cuando el cáncer comienza en las glándulas productoras del moco que protege al intestino interno

Autosoma

Uno de los 22 cromosomas no sexuales del ser humano.

Centimorgan

El centimorgan es la unidad que se usa para medir el grado de proximidad genética. Un centimorgan equivale a un 1 % de probabilidad de que un marcador en un cromosoma sea separado de un segundo marcador sobre el mismo cromosoma debido al fenómeno de entrecruzamiento (crossing-over en inglés), en una sola generación. Equivale aproximadamente, un millón de pares de bases de una secuencia de ADN en el genoma humano [tomado de: www.genome.gov]

Comunalidad

proporcion de la varianza de la i-esima variable contributiva por los m factores

Desequilibrio de Ligamento

Dos regiones (loci) presentan desequilibrio de ligamento si estas no se pueden heredar por mecanismos independientes, es decir su relación no es casual.

Error cuadratico medio

En un modelo, corresponde a la media de los errores. Es decir, si y es el valor real de una variable dependiente y $\hat{y}$ es su aproximación, entonces

$$ECM = \sum_{i=1}^{n} \frac{(y - \hat{y})^2}{n}$$

Especificidad

Corresponde a la probabilidad de diagnosticar de forma erronea a un sujeto sano (suponer que esta enfermo). En estadística, es equivalente a cometer un error de clasificación tipo I [falso positivo].

Fenotipo

Característica observable en el organismo

Haplotipo

Es una combinación de alelos de diferentes loci(región) en un cromosoma que tienden a heredarse en conjunto [36]. Es decir, un conjunto de cambios en diferentes regiones en un mismo cromosoma

Loci

Del latín, refiere un conjunto de regiones en uno de los alelos del cromosoma

Locus

Corresponde a un región de un cromosoma, es el plural es loci

Medicina de sistemas

Se refiere al estudio de enfermedades con un enfoque complejo, es decir, que las patologías se consideran entes donde interactuan una gran cantidad de factores en su desarrollo

Metástasis

Cuando el cáncer invade otras partes del cuerpo

Monocigoto

el gen dañado se replica en la generación siguiente

Mutaciones de alta penetrancia

Si una persona porta el gen asociado a la enfermedad, con alta probabilidad desarrollará la enfermedad asociada a ese fenotipo. Algunos ejemplos son el gen ATP asociado a la poliposis adenosa familiar y los genes MSH2 y MLH1 asociados a un 90 % de los casos del síndrome de Lynch

Neoplasia

Formación anormal en alguna parte del cuerpo de un tejido nuevo de carácter tumoral, benigno o maligno

Par de Bases

Un par de bases es un par de bases químicas que interaccionan entre ellas. Podemos imaginar que la doble hélice de ADN es como una escalera de mano, donde los pasamanos son las dos hebras enrolladas entre sí. La unión entre los pares de bases corresponde al peldaño de la escalera. Cada hebra está formada por la alternancia de un azúcar (desoxirribosa) y un grupo fosfato. En cada azúcar, hay anclada una de las cuatro bases nitrogenadas: adenina (A), citosina (C), guanina (G) o timina (T). Las dos hebras se mantienen juntas gracias a los puentes de hidrógeno entre las bases complementarias, es decir, la adenina con la timina, y la citosina con la guanina. tomado de [www.genome.gov]

Polimorfismo

Corresponde cambios en la secuencia del ADN producto de errores en los mecanismos de reparación y replicación, se considera un polimorfismo cuando la frecuencia en uno de sus alelos en la población es superior al 0.1 %

Susceptibilidad

Corresponde a la probabilidad de diagnosticar de forma erronea a un sujeto enfermo (suponer que esta sano). En estadística, es equivalente a cometer un error de clasificación tipo II [falso negativo].

TagSNP's

Es una representación de un grupo SNP's con fuerte desequilibrio de ligamiento en una región el genoma. Un tagsnp puede servir como representante para caracterizar al grupo de SNP [36]

Acrónimos

CCR

Cáncer Colo-Rectal

CHIBCHA

Genetic study of Common Hereditary Bowel Cancers in Hispania and the Americas

DL

Desequilibrio de Ligamento

FDR

False Discovery Rate

GWAS

Genome-wide association study

HAPMAP

Proyecto de Mapeo de Haplotipos

IARC

Centro Internacional de Registro del Cáncer

IMC

Indice de Masa Corporal

MCA

Análisis de correspondencia múltiple

OMS

Organización Mundial de la Salud

PCA

Análisis de Componentes Principales

PGH

Proyecto Genoma Humano

PLS

Mínimos cuadrados parciales

ROC

Recepción operativa de características

SNP

polimorfismos de nucleótido simple

Bibliografía

[1] Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons. 17, 18, 24, 30

[2] AL, F., GA, C., CS, F., and KM, K. (2000). Cost-effectiveness of screening for colorectal cancer in the general population. *JAMA*, 284(15):1954–1961. 1

[3] Caratachea, M. A. C. (2007). Polimorfismos genéticos: Importancia y aplicaciones. *Revista del Instituto Nacional de Enfermedades Respiratorias*, 20(3):213–221. 6

[4] Chías, A. (2016). El cáncer de colon es el 4to más frecuente en mexico. 1

[5] Chen, J., Yu, K., Hsing, A., and Therneau, T. M. (2007). A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(3):238–251. 1, 20

[6] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM. 20

[7] Consortium, . G. P. et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68. 8

[8] Cuadras, C. M. (2007). *Nuevos métodos de análisis multivariante*. CMC Editions Barcelona. 28, 49

[9] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer. 17, 19

[10] Divgi, D. R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 44(2):169–172. 27, 38

[11] Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press. 17

[12] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons. 17, 19, 22, 23, 61

[13] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874. 25

[14] Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767. 13

[15] Fernandez, M. E., Wippold, R., Torres-Vigil, I., Byrd, T., Freeberg, D., Bains, Y., Guajardo, J., Coughlin, S. S., and Vernon, S. W. (2008). Colorectal cancer screening among latinos from us cities along the texas–mexico border. *Cancer Causes & Control*, 19(2):195–206. 2

[16] Fernando Medina, M. G. (2007). *mputación de datos: teoría y práctica (Vol. 54)*. Comisión Económica para América Latina y el Caribe (CEPAL). 31

[17] Freiberg Hoffmann, A., Stover, J. B., de la Iglesia, G., and Fernández Liporace, M. (2013). Correlaciones policóricas y tetracóricas en estudios factoriales exploratorios y confirmatorios. *Ciencias Psicológicas*, 7(2):151–164. 1, 26, 38

[18] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York. 17, 18, 61

[19] Friedman, J., Hastie, T., and Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4). 65

[20] Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229. 43

[21] García-Osogobio, S., Téllez-Ávila, F. I., Méndez, N., and Uribe-Esquivel, M. (2015). Results of the first program of colorectal cancer screening in mexico. *Endoscopia*, 27(2):59–63. 16

[22] González, J. R., Armengol, L., Solé, X., Guinó, E., Mercader, J. M., Estivill, X., and Moreno, V. (2007). Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):654–655. 45

[23] Husson, F., Lê, S., and Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. Chapman and Hall/CRC. 49

[24] Iniesta, R., Guinó, E., and Moreno, V. (2005). Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos. *Gaceta Sanitaria*, 19(4):333–341. 1, 7, 45, 46

[25] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated. 17, 27, 39

[26] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90. 1

[27] Johnson, I. and Lund, E. (2007). Nutrition, obesity and colorectal cancer. *Alimentary pharmacology & therapeutics*, 26(2):161–181. 10, 11, 12

[28] Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31. 31, 32, 43

[29] Kassambara, A. (2017). *Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, volume 2. STHDA. 52, 54

[30] Kirk, D. B. (1973). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, 38(2):259–268. 38

[31] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26. 61

[32] Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18. 49

[33] Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from sweden, denmark, and finland. *New England journal of medicine*, 343(2):78–85. 1, 8, 10, 13, 86

[34] Low, S.-K., Takahashi, A., Ashikawa, K., Inazawa, J., Miki, Y., Kubo, M., Nakamura, Y., and Katagiri, T. (2013). Genome-wide association study of breast cancer in the japanese population. 8:e76463. 7, 8

[35] Luna, D. F. B., Manrique, M. A., García, M. Á. C., Corona, T. P., Velázquez, N. N. H., Espinoza, Y. M. E., Urrutia, J. M. G., Macías, E. J. R., Ramírez, G. M., Cisneros, A. A., et al. (2016). Epidemiología del cáncer colorrectal en menores de 50 años en el hospital juárez de méxico. *Endoscopia*, 28(4):160–165. 2, 16

[36] Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama*, 299(11):1335–1344. 6, 7, 89, 90

[37] Pértegas Díaz, S. and Pita Fernández, S. (2002). Cálculo del tamaño muestral en estudios de casos y controles. *Cad Aten Primaria*, 9:148–50. 30

[38] Pointet, A.-L. and Taieb, J. (2017). Cáncer de colon. *EMC - Tratado de Medicina*, 21(1):1 – 7. 1, 8, 11

[39] Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.8.10. 38

[40] Rúa, Katherine Andrea Palacio Peña, C. M. M. (2012). Bases moleculares del cáncer colorrectal. *Iatreia*, 25(2):137–148. 1, 13, 15

[41] Sánchez Cantalejo, E. (2000). Regresión logística en salud pública. *Escuela Andaluza de Salud Pública. Granada*. 29

[42] Secretaria de Salud, M. (2009). Guía de práctica clínica, detección oportuna y diagnóstico de cáncer de colon y recto no hereditario en adultos en primero, segundo y tercer nivel de atención. 1, 12, 13, 16

[43] Shirai, S., Kudo, M., and Nakamura, A. (2009). Comparison of bagging and boosting algorithms on sample and feature weighting. In *International Workshop on Multiple Classifier Systems*, pages 22–31. Springer. 20

[44] Society, A. C. (2015). Datos y estadísticas sobre el cáncer entre los hispanos/latinos 2015-2017. *Atlanta: American Cancer Society*. 9, 11

[45] Stewart, B., Wild, C. P., et al. (2017). World cancer report 2014. 1, 2, 9, 11, 12, 13, 14

[46] Telgarsky, M. and Vattani, A. (2010). Hartigan's method: k-means clustering without voronoi. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 820–827. 36

[47] Therneau, T. M., Atkinson, E. J., et al. (1997). An introduction to recursive partitioning using the rpart routines. 21

[48] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67. 31, 32, 35

[49] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. 39

[50] Winawer, S. J., Fletcher, R. H., Miller, L., Godlee, F., Stolar, M., Mulrow, C., Woolf, S., Glick, S., Ganiats, T., Bond, J., et al. (1997). Colorectal cancer screening: clinical guidelines and rationale. *Gastroenterology*, 112(2):594–642. 1, 10, 13

[51] Worthley, D. L., Whitehall, V. L., Spring, K. J., and Leggett, B. A. (2007). Colorectal carcinogenesis: road maps to cancer. *World journal of gastroenterology: WJG*, 13(28):3784. 1, 13, 14