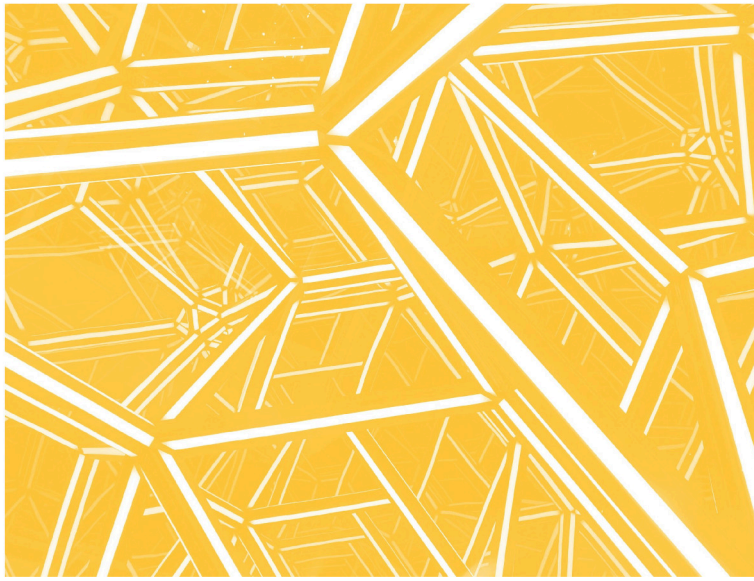
ANÁLISIS DE DATOS EN CIENCIAS DE LA SALUD MEDIANTE EL APRENDIZAJE DE IA



Vanesa Verónica Miana

COLECCIÓN UAI – INVESTIGACIÓN

UAI EDITORIAL

teseo

ANÁLISIS DE DATOS EN CIENCIAS DE LA SALUD MEDIANTE EL APRENDIZAJE DE IA

Vanesa Verónica Miana

Análisis de datos en ciencias de la salud mediante el aprendizaje de IA

Colección UAI – Investigación

UAI EDITORIAL

teseo

Miana, Vanesa Verónica
Análisis de datos en ciencias de la salud mediante el aprendizaje de IA / Vanesa Verónica Miana. – 1a ed. – Ciudad Autónoma de Buenos Aires: Teseo; Ciudad Autónoma de Buenos Aires: Universidad Abierta Interamericana, 2025. 194 p.; 20 x 13 cm.
ISBN 978-987-723-462-6
1. Cáncer. 2. Inteligencia Artificial. 3. Salud. I. Título.
CDD 006.3

© UAI, Editorial, 2025
© Editorial Teseo, 2024
Teseo – UAI. Colección UAI – Investigación
Buenos Aires, Argentina
Editorial Teseo
Hecho el depósito que previene la ley 11.723
Para sugerencias o comentarios acerca del contenido de esta obra, escríbanos a: **info@editorialteseo.com**
www.editorialteseo.com

DOI: 10.55778/ts877234626

Las opiniones y los contenidos incluidos en esta publicación son responsabilidad exclusiva del/los autor/es.

EBOOK

TeseoPress Design (www.teseopress.com)

Autoridades

Rector Emérito: Prof. Dr. Edgardo Néstor De Vincenzi
Rector: Dr. Rodolfo De Vincenzi
Vice-Rectora Académica: Dra. Ariana De Vincenzi
Vice-Rector de Gestión y Evaluación:
Ph.D., Prof. Marcelo De Vincenzi Zemborain
Vice-Rector de Investigación: Dr. Mario Lattuada
Vice-Rector de Extensión Universitaria:
Dr. Fernando Grosso
Vice-Rector de Administración: Dr. Alfredo Fernández
Decano Facultad de Medicina y Ciencias de la Salud:
Dr. Alejandro Botbol

Comité editorial

Lic. Juan Fernando ADROVER
Arq. Carlos BOZZOLI
Mg. Osvaldo BARSKY
Dr. Marcos CÓRDOBA
Mg. Roberto CHERJOVSKY
Dra. Ariana DE VINCENZI
Dr. Roberto FERNÁNDEZ
Dr. Fernando GROSSO
Dr. Mario LATTUADA
Dra. Claudia PONS
Dr. Alejandro BOTBOL

Los contenidos de libros de esta colección cuentan con evaluación académica previa a su publicación.

Presentación

La Universidad Abierta Interamericana ha planteado desde su fundación en el año 1995 una filosofía institucional en la que la enseñanza de nivel superior se encuentra integrada estrechamente con actividades de extensión y compromiso con la comunidad, y con la generación de conocimientos que contribuyan al desarrollo de la sociedad, en un marco de apertura y pluralismo de ideas.

En este escenario, la Universidad ha decidido emprender junto a la editorial Teseo una política de publicación de libros con el fin de promover la difusión de los resultados de investigación de los trabajos realizados por sus docentes e investigadores y, a través de ellos, contribuir al debate académico y al tratamiento de problemas relevantes y actuales.

La *colección investigación* TESEO – UAI abarca las distintas áreas del conocimiento, acorde a la diversidad de carreras de grado y posgrado dictadas por la institución académica en sus diferentes sedes territoriales y a partir de sus líneas estratégicas de investigación, que se extiende desde las ciencias médicas y de la salud, pasando por la tecnología informática, hasta las ciencias sociales y humanidades.

El modelo o formato de publicación y difusión elegido para esta colección merece ser destacado por posibilitar un acceso universal a sus contenidos. Además de la modalidad tradicional impresa comercializada en librerías seleccionadas y por nuevos sistemas globales de impresión y envío pago por demanda en distintos continentes, la UAI adhiere a la red internacional de acceso abierto para el conocimiento científico y a lo dispuesto por la Ley n°: 26.899 sobre *Repositorios digitales*

institucionales de acceso abierto en ciencia y tecnología, sancionada por el Honorable Congreso de la Nación Argentina el 13 de noviembre de 2013, poniendo a disposición del público en forma libre y gratuita la versión digital de sus producciones en el sitio web de la Universidad.

Con esta iniciativa la Universidad Abierta Interamericana ratifica su compromiso con una educación superior que busca en forma constante mejorar su calidad y contribuir al desarrollo de la comunidad nacional e internacional en la que se encuentra inserta.

Dra. Ariadna Guaglianone
Secretaría de Investigación
Universidad Abierta Interamericana

Índice

Prólogo 15
Dr. Ing. Sergio F. Yapur
Agradecimientos 17
Introducción 19
1. Cáncer, tóxicos ambientales y nutricionales 21
1.1. Factores predisponentes de cáncer 25
2. Aprendizaje automático (*machine learning*) 45
2.1. Algoritmos 48
3. Materiales y métodos 61
3.1. Análisis de los resultados 63
4. Resultados de aprendizaje no supervisado 67
Primer análisis 68
Segundo análisis 97
5. A modo de cierre 125
Bibliografía 135
Anexos 159
Anexo I. Codificación de las variables 159
Anexo II. Variables del análisis 1 y 2 169
Anexo III. Aprendizaje no supervisado 181
Glosario 191

Prólogo

Dr. Ing. Sergio F. Yapur

Un futuro en el cual las enfermedades se puedan prevenir ya no es un ejercicio de la imaginación, estamos frente a una nueva era de la medicina. Esto se manifiesta en la presente investigación independiente e integral de los factores predisponentes a la enfermedad neoplásica. En particular, se enfoca el análisis sobre ocho tipos de cánceres: colon, pulmón, piel, mama, cuello uterino, próstata, linfoma y leucemia. Por primera vez en nuestro país, se lleva a cabo un estudio exhaustivo y a gran escala de los factores de riesgo de cáncer, lo que permite comprender mejor la enfermedad y desarrollar estrategias de prevención efectivas.

El análisis se basa en el aprendizaje maquinal, tanto supervisado como no supervisado, que posiciona a nuestro país a la vanguardia de la oncología. Se consideran múltiples factores de riesgo comunes a diferentes tipos de cáncer, lo que abre las vías para la prevención y la detección temprana de la enfermedad. Además, se estudian las interacciones que tienen entre sí estos factores, aspecto esencial para contemplar la compleja fenomenología asociada a la enfermedad neoplásica. Debido al gran número de variables involucradas, es posible extraer tanto conclusiones de índole científica como también consejos, hábitos y costumbres para alcanzar una vida saludable.

Resulta fácil entusiasmarse por el impacto de la inteligencia artificial sobre la medicina actual. Como dijo Francis Bacon, el conocimiento es poder; y así es como podemos pensar que el poder predictivo de muchos algoritmos

solo se consigue a partir de un buen conocimiento de los datos y de los métodos involucrados. Sin embargo, esta suerte de apalancamiento cognitivo presenta desafíos, y es fundamental discernir una adquisición de datos bien diseñada, un preprocesamiento apropiado, la idoneidad del algoritmo implementado y la consistencia de las conclusiones con respecto a las etapas previas. Este proceso pormenorizado es uno de los aportes centrales de esta obra. Se elabora cada paso, se describe cada intento –fructífero o no– desde el trabajo de campo hasta las conclusiones. Al final, los mejores trabajos científicos son los que buscan la verdad por sobre todas las cosas, y la presente obra es uno de ellos.

Agradecimientos

Quisiera hacer una mención especial a la Dra. Martha Schawrcz, al Lic. Jorge Kamlosfsky y a la Dra. Julieta Palomeque.

A todas y cada una de las personas que contribuyeron con este trabajo, y a ti, que por la razón que sea estás leyendo esto y dedicándome tu tiempo.

Introducción

En la actualidad, la medicina y la inteligencia artificial (IA) se han conjugado con el fin de obtener mejores pronósticos y diagnósticos sobre la enfermedad neoplásica (1). Según estimaciones del Observatorio Global del Cáncer (Globocan), de la Agencia Internacional de Investigación sobre Cáncer (IARC), en Argentina ocurrieron 130.878 casos nuevos de cáncer en ambos sexos en el año 2020.[1]

Este trabajo se originó con el interés de ampliar el conocimiento sobre los factores predisponentes a la aparición de cáncer incorporando las ventajas que presenta la inteligencia artificial para el análisis de un gran número de variables. Se diseñó un modelo de investigación basado en una muestra de individuos lo suficientemente amplia que asegurase la validez de los resultados. Lo suficientemente abarcadora en variables para permitir a los algoritmos de aprendizaje automatizado establecer modelos de predicción entre hábitos y cáncer que hubieran podido pasar desapercibidos mediante el análisis de la epidemiología clásica, sea esta descriptiva, de correlación o de inferencia a través de medidas de dispersión, tablas de frecuencia, coeficientes de correlación de Pearson, regresión lineal, T de student, ANOVA, por solo nombrar algunos ejemplos (2).

El aprendizaje automatizado es una rama de la inteligencia artificial que, a través de los diferentes tipos de aprendizaje (supervisado, no supervisado, semisupervisado), realiza el procesamiento de un gran conjunto de

1 https://www.argentina.gob.ar/salud/instituto-nacional-del-cancer/estadisticas/incidencia

información en un tiempo compatible con las necesidades del análisis. A fin de conocer qué algoritmos son los más precisos para la detección de factores vinculantes a la enfermedad neoplásica, se abordó el trabajo de investigación con un enfoque interdisciplinar. Se amplió el tamaño de la muestra con el objeto de obtener mayor cantidad de datos válidos que permitiesen establecer modelos de predicción a los algoritmos de aprendizaje automático que caractericen la relación entre hábitos y enfermedad neoplásica.

La IA permite obtener mejores resultados para las predicciones y diagnósticos de enfermedades al poner mayor énfasis en el desarrollo de programas informáticos inteligentes. Los trabajos publicados hasta la fecha se fundaron en bases de datos públicos y se refieren a la importancia del uso de la inteligencia artificial para el diagnóstico y tratamiento de enfermedades. Ninguno de los trabajos mencionados hace referencia a la utilización de la IA en estudios de prevención. Moreno Luna (3) remarca la importancia de la aplicación de la IA en la medicina y en los diagnósticos precoces para el procesamiento de imágenes de lunares, con el fin de detectar melanomas; para ello, utilizó la base de datos ISIC 2016, que cuenta con novecientas imágenes en su banco de datos. También, Prado Vázquez (4) concluyó en su trabajo de clasificación molecular en capas de información sobre los datos moleculares del melanoma basados en modelos gráficos probabilísticos, logrando establecer la correlación entre la información registrada en las historias clínicas de los pacientes y los resultados obtenidos en su análisis.

1

Cáncer, tóxicos ambientales y nutricionales

En todo organismo las células se multiplican, nacen y mueren. Algunas lo hacen a grandes velocidades, mientras que otras lo hacen muy lentamente de acuerdo con las necesidades de los tejidos y organismos. El balance entre los procesos de nacimiento y muerte se mantiene mediante un complicado sistema de controles genéticos que involucra señales de crecimiento, de inhibición y de muerte celular. La mayoría de los cánceres se produce por daños genéticos en los componentes del sistema de controles. Al perder el control, las células se multiplican rápidamente, forman tumores, se malignizan.

El término cáncer engloba un conjunto de más de doscientas enfermedades de origen genético. La mayoría de las mutaciones se produce en las células somáticas y se transmite a las células hijas durante la mitosis. Solamente cuando la mutación se produce en el ADN de los gametos, la probabilidad de desarrollar un cáncer se hereda de padres a hijos (5, 6).

El cuerpo humano está formado por alrededor de 10^{14} células que se multiplican por división celular, siendo el daño o el envejecimiento la señal de que deben ser reemplazadas por células nuevas. En el normal funcionamiento del ciclo celular intervienen tres tipos de genes. Estos son los protooncogenes que participan en la formación y multiplicación normal de las células. Incluyen genes que codifican para factores de

crecimiento, receptores de factores de crecimiento y todas las proteínas que intervienen en la transducción de señales de los factores de crecimiento. El protooncogén mutado se denomina oncogén, y la mutación da como resultado un aumento de la velocidad de la división celular, teniendo características autosómicas dominantes. Basta que un alelo mute para que la mutación se exprese fenotípicamente. Otros son los genes supresores de tumores, que inhiben el crecimiento celular. Pueden sufrir cambios en su ADN que disminuyan esta función, en cuyo caso las células tumorales se multiplican sin control. Se manifiestan con características autosómicas recesivas, es decir, ambos alelos tienen que estar mutados para que se expresen fenotípicamente. Entre los ejemplos de genes supresores que codifican proteínas se encuentran: a) p53 (gen que codifica la proteína tumoral 53) y b) Bcl-2 (gen que codifica la proteína 2 de la leucemia/linfoma de células B) (7). Por último, están los genes de reparación del ADN, y existen distintos mecanismos para tal fin.

Cuando los mecanismos de reparación no son suficientes o efectivos, se produce una mutación que se manifiesta como un cambio permanente en el genoma. Los destinos de una célula que ha sufrido daño genético primario son: la reparación, el aumento de la inestabilidad y la senescencia o la apoptosis; esta última funciona como un mecanismo de protección en el cual se sacrifican las células dañadas, pero el organismo prevalece.

Los sistemas de control del ciclo celular responden a información del proceso al que están controlando mediante sensores. Así, detectan que se haya completado exitosamente la replicación del ADN o que los cromosomas resultantes se hayan separado correctamente.

Si por algún motivo esto no sucediera, se envían señales para retardar la progresión a la fase siguiente y para permitir que los sistemas de reparación del genoma actúen. Los puntos de control son monitoreados por proteínas inhibidoras del ciclo celular. Estas proteínas son codificadas por genes supresores de tumores, ya que su acción inhibe la división celular y promueve la reparación o la apoptosis de la célula dañada (7-9).

Las mutaciones que dan origen a células tumorales se producen con mucho mayor frecuencia en las células somáticas, por lo que no son transmisibles de progenitores a su descendencia, pero sí a las células hijas, lo que da lugar a un clon de células alteradas. Una célula tumoral se caracteriza por la pérdida de la inhibición por contacto, el aumento en la velocidad de proliferación celular, la pérdida de diferenciación (asociada a cambios metabólicos) y la formación de masas tumorales. A estas características se suman al fenotipo tumoral: inestabilidad genómica, invasión de los tejidos adyacentes, aparición de subpoblaciones migratorias, capacidad de proliferar en otros tejidos y formar metástasis, neoangiogénesis e inmortalidad celular.

El cáncer de cada persona es una combinación única de cambios genéticos. Es un proceso microevolutivo que, en general, se produce en una escala de tiempo de años, se necesitan no menos de cuatro mutaciones activantes de protooncogenes e inhibitorios de genes supresores de tumores en la misma célula para que un tumor comience a malignizarse. Dentro del mismo tumor es posible que las diversas células tengan cambios genéticos distintos (9, 10).

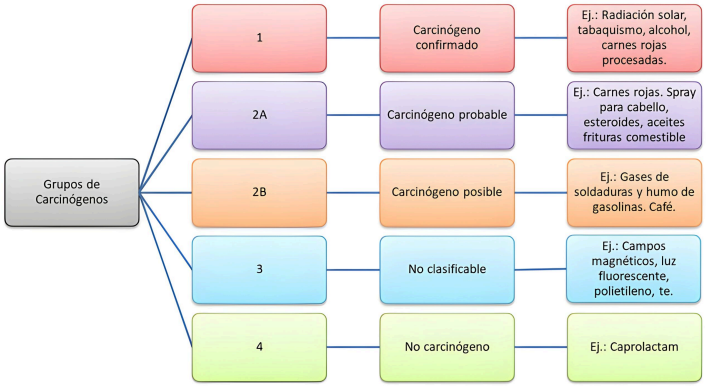
La carcinogénesis se divide en tres etapas: iniciación, promoción y progresión. La primera etapa es un cambio celular irreversible o mutación del ácido

desoxirribonucleico que, si bien es compatible con la vida de la célula, involucra cambios en el sistema de señales bioquímicas asociadas al control de la proliferación celular o de la intercomunicación con otras células. En la segunda etapa, la célula dañada recibe estímulos para su crecimiento y proliferación y comienza a dividirse sin control. En la última, las células de las subpoblaciones que han sido generadas a partir de la expansión clonal inicial han establecido su malignidad y la enfermedad avanza (11).

Si bien el origen del cáncer es genético, la tumorigénesis surge de una interacción entre la genética y el medio ambiente. El cáncer se origina principalmente como consecuencia de la exposición prolongada a carcinógenos, entendidos como aquellos agentes químicos, físicos o biológicos que aumentan la tasa de mutación en células y tejidos específicos, y que son considerados carcinogénicos (9). Las mutaciones en el ADN pueden originarse por diversas causas, tales como los radicales libres del oxígeno (O_2), errores en la división celular, herencia, factores ambientales (exógenos y endógenos) y algunos virus (11).

La IARC (Agencia Internacional para la Investigación del Cáncer) clasificó los carcinógenos en cuatro grupos, según las pruebas científicas existentes sobre la carcinogénesis de compuestos y factores físicos (12,13), como se puede apreciar en la siguiente figura.

Figura 1. Clasificación de los carcinógenos



Fuente: elaboración propia con base en la referencia de la IARC, OMS.

La interacción de varios factores de riesgo contribuye al desarrollo de la enfermedad neoplásica, tal es el caso de a) factores ambientales exógenos y endógenos, y b) los factores individuales (como, por ejemplo, la predisposición genética). A través de estudios epidemiológicos, se ha demostrado que entre el 80 % y el 90 % de los tumores malignos tienen su causa en carcinógenos derivados del comportamiento humano (14).

1.1. Factores predisponentes de cáncer

Dentro del grupo de los factores predisponentes físicos se encuentran la exposición a campos electromagnéticos, radiación ionizante y radiación ultravioleta. En cuanto a los factores químicos, abarcan el alcohol, el tabaco, los compuestos organoclorados y los hidrocarburos aromáticos policíclicos (HAP). Por otro lado, los factores biológicos comprenden el sobrepeso

y la obesidad, la dieta, los compuestos mutagénicos y cancerígenos presentes en los alimentos, el ejercicio físico y las infecciones.

Uno de los factores predisponentes físicos relacionados con la exposición es la presencia de campos electromagnéticos. Estos se definen como la existencia de diferencias de voltaje que, al principio, generan un campo eléctrico, el cual puede existir incluso sin corriente. Asimismo, las corrientes eléctricas son las responsables de originar los campos magnéticos. La magnitud de estos últimos varía en función del consumo de energía, mientras que la fuerza del campo eléctrico se mantiene constante.

En la vida diaria, se utilizan equipamientos que generan campos magnéticos, cuyas potencias y efectos sobre el organismo aún se desconocen. La bibliografía define que un campo magnético de 50/60 Hz puede ser un probable factor tumoral. Ejemplos de fuentes comunes incluyen ordenadores (240 Hz), televisores (200 Hz, 800 Hz, 1000 Hz), teléfonos móviles e inalámbricos (90 Hz, 120 Hz), y tecnologías como wifi o Bluetooth (2472 MHz) (11).

Existen estudios epidemiológicos que indican una relación entre la exposición a un campo magnético de baja frecuencia (entre 0 Hz y 300 Hz) y el riesgo de padecer cáncer. Se han encontrado resultados significativos que vinculan este factor a los diagnósticos de gliomas, meningiomas y leucemias (15).

La radiación ionizante también se encuentra entre los factores predisponentes físicos y se la entiende como la emisión de alta energía con la fuerza necesaria para separar un electrón de un átomo o de una molécula y causar su ionización. Se mide en Sievert (Sv), que es la unidad del Sistema Internacional de Unidades (SI) para cuantificar el efecto biológico del depósito de un joule de energía de radiación en un kilogramo de tejido humano (16, 17).

Las fuentes de esta radiación pueden ser naturales, como los rayos cósmicos que atraviesan la atmósfera terrestre, o artificiales, como los equipos de imágenes médicas: radiografía, tomografía computarizada (TC) o tomografía por emisión de positrones (PET) (18).

En protección radiológica se habla de una milésima parte, es decir, en mili Sieverts (mSv). El límite de exposición que puede recibir un individuo es de 1 mSv como dosis total por año. Ejemplos de estas radiaciones son la radiografía de tórax: 0.02 mSv, la mamografía: 0.60 mSv (17). El riesgo de cáncer se incrementa en dosis de 0.5 Sv que se detectan a lo largo de los años y tienen como variables dependientes el sexo, la edad y el tipo de neoplasia que se desarrolla (19, 20).

La inducción del cáncer por medio de radiaciones ionizantes se produce por activación de los protooncogenes o por la inactivación de las células supresoras de tumores. También, por acción directa sobre el ADN o de forma indirecta por los radicales libres generados por la interacción de la radiación con el agua de la célula (21).

Los estudios diagnósticos que utilizan rayos X no implican dosis de radiación extremadamente elevadas. Se ha demostrado, epidemiológicamente, que dosis superiores a la franja entre los 50 mSv y los 100 mSv o exposiciones agudas entre 10 mSv y 50 mSv incrementan el riesgo de cáncer (17, 19, 20).

Es difícil estimar y cuantificar el daño que pueden producir dosis inferiores a 10 mSv, pero esto no significa que el riesgo sea inexistente. Este tipo de radiación puede dañar el ADN, lo que ocasionaría la formación tumoral. Así, aunque el riesgo sea bajo, si afecta a un gran número de individuos, puede resultar en un problema importante de salud pública (22, 23).

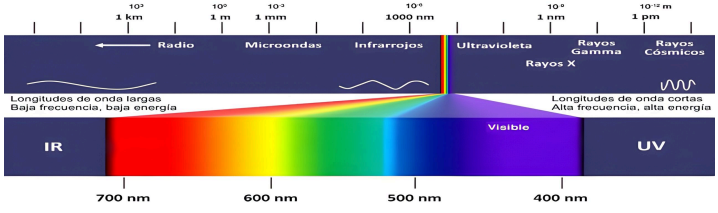
Por último, en este grupo de factores físicos se encuentra la radiación ultravioleta, definida como la exposición crónica y excesiva a la luz solar. Esta conduce a efectos adversos prematuramente, tales como eritema o quemaduras solares,

envejecimiento acelerado de la piel y carcinogénesis postsolar. Se le atribuyen estos daños a la luz ultravioleta procedente del cosmos (24).

La luz se define como ondas cortas o una radiación electromagnética. Una onda corta es entendida como una perturbación o una magnitud física que varía y que se propaga en el espacio y durante el tiempo. El espectro electromagnético abarca desde las ondas de radio con longitudes de onda de kilómetros e inclusive los rayos gamma (por debajo de 0,1 nm). La luz visible es una pequeña parte del espectro.

La radiación ultravioleta, o UV, tiene una longitud de onda mayor a la de los rayos X y, a su vez, más corta que la onda de la luz visible. Estas ondas son las que implican un mayor efecto sobre las células al provocar lesiones en el ADN; debido a que su longitud de onda corta es inferior a la emitida por la luz visible. Se clasifican en tres tipos sobre la base de dicha longitud, tal como se puede observar en la gráfica siguiente (25).

Figura 2. Longitudes de onda: UV-A 320-400 nm, UV-B 280-315 nm y UB-C 200-280 nm[2]



Fuente: elaboración propia.

Los efectos de esta radiación sobre la piel se miden en unidades de Dosis Estándar de Eritema (DES), las cuales equivalen a 100 J/m^2 (26). La radiación UV-A es la longitud

2 Fuente: https://tinyurl.com/mwajjpav

de onda más cercana a la de la luz visible y sí llega a la superficie terrestre. Es capaz de generar la aparición de ERO (especies reactivas de oxígeno), como el oxígeno singlete, y favorece las lesiones como la 8-oxoguanina, que es altamente mutagénica. La exposición a esta radiación puede ser a causa del uso de lámparas en el ámbito laboral, y puede traer como consecuencia problemas oculares y cutáneos en humanos, porque al igual que la radiación UV-B (ultravioleta de onda corta) daña el ADN, de modo que se lo considera carcinógeno (27, 28).

La exposición a la radiación UV-B induce la aparición de mutaciones, así como la inmunosupresión local en la piel; también, daña las vías que permiten la inducción a la apoptosis, y promueve la proliferación de células alteradas e inmaduras. Los rayos UV-C (rayos ultravioleta de onda muy corta) tienen una longitud de onda entre 200 nm y 280 nm, por lo que son potentes y producen mucho daño a la superficie que resulta afectada.

La luz ultravioleta no solo puede ser emitida de manera externa, sino también producida artificialmente. Ejemplos de estas emisiones son las generadas por ciertas lámparas LED y, principalmente, por las camas solares, que tienen efectos cancerígenos.

Las dosis de radiación UV-A emitidas por las camas de bronceado superan los límites tolerados por la piel. Estudios sobre esta radiación UV artificial han demostrado que el uso frecuente y prolongado de solarios aumenta el riesgo de desarrollar melanoma y carcinoma escamoso de piel (27-30).

Otro grupo de factores predisponentes estudiados son los químicos, entre los que se incluyen el alcohol, el tabaco, los compuestos organoclorados y los hidrocarburos aromáticos policíclicos (HAP).

El alcohol se refiere a las bebidas alcohólicas que contienen una mezcla de diferentes sustancias, siendo el etanol la más abundante. Este compuesto es un líquido claro, incoloro, volátil e inflamable que es metabolizado por el hígado a una velocidad aproximada de entre 10 g y 15 g por hora. Su valor energético es de 7,1 kcal por gramo de alcohol.

Los mecanismos por los cuales las bebidas alcohólicas pueden ser carcinogénicas aún no están completamente determinados. Sin embargo, se atribuyen efectos genotóxicos al metabolito acetaldehído. Esta sustancia se ha establecido como un factor de riesgo para el desarrollo de carcinoma hepatocelular, carcinoma de células escamosas de esófago y cánceres de cavidad oral, faringe, laringe, colon-recto y mama (en mujeres) (31-35).

El segundo factor descripto es el tabaquismo, definido como una enfermedad adictiva crónica. La nicotina es el componente químico responsable de la adicción. Representa una de las mayores amenazas para la salud a nivel mundial. Anualmente, fallecen por esta causa unos siete millones de tabaquistas directos y un millón doscientos mil fumadores pasivos (36, 37). El tabaco sufre una combustión incompleta durante su consumo, que produce siete mil sustancias químicas, de las cuales cincuenta son cancerígenas (38, 39), entre ellas el benzopireno, que es el compuesto de mayor relevancia y un indicador principal de exposición (40, 41).

El riesgo también es mayor para los fumadores que para los exfumadores (42), siendo este hábito no solo responsable de la incidencia de cánceres como el de pulmón, sino que también eleva la mortalidad de cáncer de colon, recto y próstata (43). El hábito de fumar actúa como un carcinógeno vinculado con el virus del papiloma humano (VPH o HPV), que produce cáncer de cuello de útero (44, 45). El tabaquismo tiene una relación causal al menos con dieciséis tipos de cáncer, clasificados como tumores hematopoyéticos de laringe, riñón,

vejiga, páncreas, esófago, faringe, estómago, hígado, cavidad nasal, cuello uterino, carcinoma escamoso bucal y broncopulmonar. A su vez, la evidencia demuestra que fumar puede causar cáncer de mama y leucemia en niños (35, 46).

El fumador pasivo aspira casi el 20 % del humo ambiental que exhala un fumador activo. Este humo tiene mayores concentraciones de partículas que se mantienen en suspensión, son preferentemente absorbidas por la vía aérea superior y la cavidad bucal de aquellos sujetos que estén expuestos (39, 43, 44).

Como tercer factor químico se encuentran los compuestos organoclorados, sustancias con propiedades altamente tóxicas, contaminantes ambientales permanentes y que se presentan en bajas dosis. Debido a la alta afinidad por las grasas y su baja solubilidad en el agua, se las encuentra en la grasa de las diferentes especies de animales. Las dioxinas (2, 3, 7, 8-tetraclorodibenzo-paradioxina-TCDD) y de los policlorobifenilos (PCB) (desde el 2, 3, 4, 7, 8-pentaclorodibenzofurano al PCB 126) se pueden encontrar en la carne, los lácteos, los huevos y el pescado (47-49).

Por último, se encuentran los hidrocarburos aromáticos policíclicos (HAP), un grupo de compuestos químicos muy amplio. Comúnmente se presentan en mezclas complejas, son muy abundantes en la naturaleza y son el resultado de la utilización de agroquímicos y de una combustión de hidrocarburos que reaccionan de manera incompleta con el oxígeno, lo que da como producto final monóxido de carbono y HAP.

La dieta es una fuente de exposición a estos hidrocarburos a través del consumo de alimentos cultivados en zonas contaminadas, donde los HAP están presentes en los alimentos, debido a la pirólisis (combustión incompleta) de carbohidratos, aminoácidos y ácidos grasos. El ejemplo más claro de esto sucede con el proceso de cocción de

carnes a fuego directo (alta temperatura) o en los procesos de ahumado, donde se produce la formación de HAP sobre la superficie de estas (49, 50).

La medición de dichos policíclicos en diferentes estudios ha permitido establecer que la concentración aproximada que puede encontrarse en un alimento incluye entre quince y veinte tipos diferentes, siendo estas las fuentes más frecuentes de exposición en la población humana (51). Dentro de los más representativos se encuentran: acenafteno, acenaftileno, antraceno, benzo[a]antraceno; benzo[a]pireno, benzo[e]pireno, benzo[b]fluoranteno, benzo[g,h,i]perileno, benzo[j]fluoranteno, benzo[k]fluoranteno, criseno, dibenzo[a,h]antraceno, fluoranteno, fluoreno, indeno[1,2,3-c,d]pireno, fenantreno, pireno.

El benzo[a]pireno (BaP) es el más estudiado; existen límites máximos tolerables para los distintos alimentos y los procesos a los que son sometidos (1 % al 15 % del total de HAP) (52, 53). En estudios experimentales con BaP, se ha comprobado que produce mutaciones en genes supresores de tumores, como el gen p53 (54-57). Así ocurre también con el alquitrán de hulla, un subproducto que se obtiene del carbón mineral, cuando este se calienta a 1000 °C mediante pirólisis. Su uso está destinado a la producción de sustancias químicas refinadas, como el benceno, tolueno, xileno (dimetilbenceno), naftaleno y una gran variedad de compuestos aromáticos (58). El contacto con este hidrocarburo aromático y sus residuos puede ocurrir principalmente por exposición ocupacional, y es causante de cánceres de piel, pulmón, vejiga, riñón y de las vías digestivas. Además, se lo utiliza en la preparación de productos para tratar distintas afecciones de la piel, como eczema, psoriasis y caspa (59).[3]

---

3 Aclaración: esta última mención no será objeto de estudio en este trabajo.

A su vez, el benceno es un hidrocarburo aromático obtenido de la destilación del alquitrán de hulla, que se mantiene en estado líquido a temperatura ambiente y es altamente tóxico e inflamable. Se emplea en la industria como disolvente y en la fabricación de colorantes y plásticos, entre otros usos (60). La IARC estableció la carcinogenicidad de esta sustancia al asegurar que, en el cuerpo humano, es la causa del daño al sistema hematopoyético. En estudios epidemiológicos, se confirmó la relación causal entre la exposición al benceno y el desarrollo de leucemia mielógena aguda[4] (60-62).

Existen otros compuestos, como los utilizados en la industria, y en particular el uso de derivados del petróleo es una fuente importante de contaminación ambiental. De las actividades de esta categoría pueden liberarse benceno, plomo, arsénico y HAP, por lo que estos compuestos requieren actividades reguladas y controladas en cuanto a sus niveles de seguridad en la manipulación y manejo de sustancias químicas. La bioacumulación de estos tóxicos genera en la población circundante problemas endocrinos, respiratorios, inmunológicos, renales, hepáticos, alergias y malformaciones genéticas (teratógenos), entre otras afecciones (63, 64). Li *et al.* (65) definen la epigenética como "los factores y procesos moleculares alrededor del ADN que regulan la actividad del genoma, independientemente de la secuencia del ADN, y son mitóticamente estables". Esta y la genética de los individuos, al exponerse a ambientes tóxicos o consumir alimentos con bioacumulación de ellos, afectan la salud humana.

Los compuestos halogenados son actualmente veintitrés contaminantes orgánicos persistentes, que incluyen las dioxinas y los insecticidas de diclorodifeniltricloroetano

---

4 Tampoco este punto será objeto de estudio aquí.

(DDT). Estos compuestos comparten analogía con los policlorobifenilos y la IARC los ha clasificado como carcinógenos del grupo 2A para los seres humanos. Los más estudiados son el p,p'-diclorodifenil-trichloro-etano (y su metabolito p,p'-dicloro- difenil-dicloro-etileno), el f3-hexaclorociclohexano y el hexaclorobenceno (grupo 2B-IARC), que presentan interferencia a nivel endocrino y una posible relación con el cáncer de mama y de tiroides en algunos compuestos (49, 52).

Respecto de los agroquímicos, según la bibliografía, se incluyen estudios como el de Duarte, entre otros, quien realizó una investigación en las provincias de Santa Fe, Córdoba y Entre Ríos en la cual relevó altos índices de mortalidad por cáncer en esta región de Argentina. El trabajo incluyó la correlación entre los cánceres más frecuentes en el área estudiada (pulmón, colorrectal, próstata y mama) y el trabajo agrícola e industrial (66). El auge y mantenimiento durante años del uso de los agroquímicos como organoclorados y el glifosato (N-fosfonometil) glicina tienen como característica la persistencia y la bioacumulación. Estos compuestos no tienen un proceso de descomposición rápido, lo que contribuye a su persistencia, bioacumulación y lipofilicidad. Por lo tanto, están presentes en los alimentos y contribuyen a convertirlos en factores predisponentes de enfermedades neoplásicas (64, 67).

Finalmente, la acrilamida es un hidrocarburo aromático utilizado en la industria del plástico, el papel y el tratamiento del agua potable y aguas residuales. Esta sustancia química, que es un monómero de vinilo hidrosoluble, también se encuentra en alimentos, envases de los alimentos, selladores y adhesivos. Además, forma parte del humo del tabaco.

Los alimentos que contienen carbohidratos y el aminoácido asparagina, al ser expuestos a altas temperaturas (superiores a 120 °C) y a bajas condiciones de humedad, sufren una reacción de reducción de azúcares (almidón) que provoca la descarboxilación y desaminación de la asparagina, lo que produce acrilamida como resultado final. Este proceso se conoce como reacción de Maillard. Los métodos de cocción que favorecen esta reacción son freír, asar o cocinar al horno (por ejemplo, en papas fritas, galletas, pan o café), mientras que hervir los alimentos o cocinarlos en horno microondas no genera este efecto. Un caso particular es la formación de acrilamida por pirólisis del gluten del trigo (68-70).

Los factores biológicos son los más estudiados en el ámbito de la medicina y las ciencias de la salud por su vinculación directa con el cuerpo humano. Entre ellos se incluyen el sobrepeso y la obesidad, así como la dieta, los compuestos mutagénicos y cancerígenos presentes en los alimentos, la actividad física y las infecciones.

En cuanto al sobrepeso y la obesidad, se entienden como una acumulación excesiva de grasa corporal que puede ser perjudicial para la salud. Estas condiciones se clasifican según el índice de masa corporal (IMC), y se considera sobrepeso cuando el IMC es superior a 25 y obesidad cuando supera los 30. El IMC se calcula dividiendo el peso en kilogramos por el cuadrado de la estatura en metros (IMC = peso [kg] / estatura $[m]^2$) (71).

La obesidad se ha convertido en un problema de salud pública debido a su aumento exponencial, que ha alcanzado dimensiones pandémicas. Los sujetos con un incremento significativo en la cantidad de tejido adiposo desarrollan un estado inflamatorio crónico leve, lo que conlleva un aumento a nivel sistémico de mediadores proinflamatorios estimulados por señales de origen exógeno o endógeno.

Este proceso proinflamatorio se vincula con enfermedades degenerativas crónicas, como la diabetes tipo 2, la resistencia a la insulina y las enfermedades cardíacas, así como con el daño al ADN ocasionado por el estrés oxidativo.

La inflamación provocada por el incremento del tejido adiposo favorece la infiltración y activación de macrófagos, así como el aumento en la acción de citocinas de respuesta proinflamatoria, como el factor de necrosis tumoral alfa (TNF-α), la interleucina 1-beta (IL-1β), la interleucina-12 (IL-12), el interferón-γ (IFN-γ) y la interleucina 6 (IL-6).

Además, la inflamación asociada con la obesidad favorece la tumorigénesis mediante el aumento de adipocitos hipertróficos, que estimulan la producción del factor inducible por hipoxia 1-alfa (HIF-1α) y promueven la angiogénesis (72).

En 2007, el Fondo Mundial para la Investigación del Cáncer informó que un 35 % de las incidencias de enfermedad neoplásica a nivel mundial se pueden atribuir a la obesidad y al sobrepeso. Entre los principales cánceres relacionados se encuentran el colorrectal, los cánceres de mama, de esófago, de estómago y de páncreas (73-75).

El aumento del riesgo de cáncer está más asociado a condiciones como la obesidad y a una alimentación desequilibrada, pobre en verduras y frutas, más que a la ingesta de potenciales carcinógenos en los alimentos. Sin embargo, entre los potenciales cancerígenos más importantes de la dieta se han propuesto las nitrosaminas, las aminas heterocíclicas y los hidrocarburos policíclicos aromáticos, los cuales derivan del tratamiento en la elaboración de ciertos productos alimentarios o del tipo de cocción de alimentos como las carnes (76, 77).

La dieta, relacionada con el factor predisponente anterior, es el patrón alimentario constituido por un conjunto de alimentos y productos alimentarios que una población

consume de manera habitual, sea varias veces al día, a la semana o al mes. Este reúne las representaciones, creencias, conocimientos y prácticas vinculadas a la alimentación, ya sea por tradición familiar o aprendidas y compartidas por un grupo social entre sí, que se desenvuelve dentro de una cultura a lo largo de los años.

La alimentación constituye una variable multidimensional y, por tanto, medir su exposición resulta complejo, pero importante para poder evaluar los alimentos que tienen incidencia en la enfermedad neoplásica (78, 79). A partir de ella es posible observar un estilo de vida de la cultura propia de una población que vive en un hábitat con determinadas características geográficas y climáticas que limitan, en algunos casos, la disponibilidad y acceso a los alimentos y la tecnología, producción, manufactura, procesamiento, distribución y venta. Desde el proceso de globalización, se cambiaron los patrones alimentarios en gran parte de los países y culturas con prácticas poco saludables, sumado a un ritmo de vida más vertiginoso, que facilita el consumo de alimentos procesados y bajos en nutrientes.

Los alimentos ahumados, en escabeche, con alto contenido graso o calientes en extremo, debido a su pretratamiento o en su producción para el consumo, pueden contener altas dosis de nitritos y nitratos, entre los carcinógenos más importantes. Cabe destacar que se mencionan como adyuvantes en la transformación maligna el bajo consumo de calcio y vitaminas antioxidantes, así como el alto consumo de sodio. También es conocido que la dieta con alta ingesta de carnes durante toda la vida en mujeres se asocia con un aumento del 43 % en el riesgo de cáncer de mama premenopáusico (80, 81).

En la vida cotidiana, el consumo de bebidas no alcohólicas puede traer aparejados riesgos para el desarrollo de enfermedad neoplásica sin que sea tenido en cuenta. De las infusiones más conocidas, desarrollaremos el riesgo que puede resultar de un excesivo consumo de mate y de bebidas azucaradas. Estas últimas son ricas en calorías "vacías" y pueden causar aumento de peso. Las Guías Alimentarias 2020-2025 recomiendan limitar los azúcares añadidos a menos del 10 % de su ingesta diaria de calorías. Una recomendación para las mujeres es el consumo de seis cucharaditas (100 calorías) de azúcar al día, mientras que se aconseja nueve cucharaditas (150 calorías) diarias para los hombres.

El azúcar también puede encontrarse bajo diferentes nombres, como jarabe de maíz, dextrosa, fructosa, jarabe de maíz rico en fructosa, miel, sirope, sirope de agave, jarabe de arroz marrón, melaza, jugo de caña evaporado y edulcorantes no nutritivos, como el aspartamo, acesulfame K o estevia, que no aportan calorías (82).

Los riesgos asociados a la salud están dados por el aumento de peso, la resistencia a la insulina que puede derivar en diabetes tipo 2 y el incremento a padecer enfermedades cardiovasculares. Cabe recordar que en un individuo la inflamación generada por la obesidad puede desencadenar enfermedades sistémicas crónicas (83, 84).

El mate es una infusión que se prepara a base de hojas de yerba mate que se secan, cortan y muelen. Al igual que el café y el té, tiene propiedades de antimutagénico, antioxidante y antiinflamatorio. Sin embargo, debido al tratamiento previo al que deben someterse las hojas y plantas de yerba mate, se han encontrado concentraciones de HAP (benzopireno) a causa del proceso de secado (temperaturas elevadas) y procesamiento de las hojas (corte y molienda). Se ha asociado significativamente el consumo

de mate con el cáncer de esófago. Los estudios realizados demuestran un nexo, aunque no hay evidencia consistente sobre la relación dosis-respuesta, es decir, asociación de mayor riesgo a mayor consumo. Por consiguiente, se considera que la relación entre la enfermedad neoplásica en el esófago y el mate puede tener dos vías: la presencia del cancerígeno benzopireno o un factor externo, como el agua caliente utilizada para el consumo de esta bebida, que genera una irritación crónica en la mucosa de este órgano (85-87).

En lo que se refiere a los compuestos mutagénicos y cancerígenos en los alimentos, entendemos por bacterias a los microorganismos unicelulares que se reproducen de forma exponencial por fisión binaria, siendo algunas de ellas patógenas o no, y son utilizadas en la industria alimenticia y farmacéutica para la elaboración de diferentes productos de consumo (yogures, antibióticos, entre otros) (88, 89).

Muchos de estos microorganismos procariontes y eucariontes (bacterias, virus, hongos) son responsables de la intoxicación alimentaria, y los signos, síntomas y gravedad dependerán del patógeno que la haya originado y del individuo que resulte afectado. La Administración de Alimentos y Medicamentos de los Estados Unidos (FDA, por sus siglas en inglés) informa sobre las hospitalizaciones y muertes por estas causas. Los patógenos más mencionados son: a) *Campylobacter jejuni*, b) *Clostridium perfringens*, c) *Clostridium botulinum*, d) *Escherichia coli 0157:H7*, e) *Listeria monocytogenes*, f) virus de la hepatitis A y E, g) *Bacillus Cereus*, h) *Shigella*, i) *Staphylococcus aureus*, j) *Vibrios* (vibriosis) (90, 91).

Algunos de los compuestos mencionados, ya sean naturales o agregados por el hombre, no causan una enfermedad aguda cuando se consumen en pequeñas

cantidades. Sin embargo, se acumulan en el organismo con el tiempo, lo que puede derivar en graves problemas de salud a largo plazo. Un ejemplo de ello son las micotoxinas, que producen enfermedades conocidas con el nombre de micotoxicosis. A bajos niveles, estos son carcinogénicos, mutagénicos, teratogénicos; además, producen alteraciones mitóticas, interfieren con el desarrollo fetal, pueden reducir la tasa de crecimiento y alteran la respuesta inmunológica, lo cual aumenta la susceptibilidad a infecciones. También afectan diversos órganos y sistemas, como el hígado, los riñones y los sistemas nervioso, endocrino e inmune.

Las cepas toxigénicas de hongos producen aflatoxinas, compuestos que se incluyen dentro de las micotoxinas. Estas sustancias químicas son producidas por algunas variantes del hongo filamentoso hialino y saprofito *Aspergillus*, como: *Aspergillus flavus* (B1) y *Aspergillus parasiticu*s, *Penicillium* y *Fusarium*. Su toxicidad para humanos y ganado puede causar enfermedad y muerte. Estas toxinas se encuentran en alimentos como maíz, arroz, café, leche, legumbres, entre otros, especialmente cuando no se han tratado adecuadamente en su proceso de cosecha y comercialización. Aunque los hongos pueden ser destruidos durante los procesos industriales o de cocción, las toxinas que generan persisten en los alimentos (92, 93).

Las aflatoxinas que tienen implicancias para la salud humana son las producidas por especies de *Aspergillus flavus*, que generan las aflatoxinas Bl y B2 (AFBl y AFB2, respectivamente). La otra especie, el *Aspergillus parasiticus*, produce las aflotoxinas B, Gl y G2, Ml y M2, metabolitos de AFBl y AFB2 (pueden encontrarse en la leche). Las especies *Aspergillus ochraceus* y *Penicillium verrucosum* producen ocrotoxina A y diversas especies de *Fusarium*

que infectan el maíz, producen fumonisinas Bl, B2 y C4 (son micotoxinas producidas por *Fusarium verticillioides* y *F. proliferatum,* que contaminan los cereales).

La aflatoxina B1 es el factor que impide el desarrollo fetal normal y tiene la mayor capacidad de carcinogénesis, dado que acelera la enfermedad. Además, este compuesto puede provocar alteraciones genéticas repentinas y permanentes, incluyendo una mutación específica en el codón 249 del gen supresor p53, lo cual está directamente relacionado con la génesis de tumores (92, 94, 95). Existen más de cuatrocientas micotoxinas, de las cuales las aflatoxinas, las ocratoxinas, las fumonisinas y la cearalenona tienen injerencia en la inducción de los cánceres de mama, hígado, esófago y próstata (96).

Es plausible que, durante el almacenamiento de granos como el maíz y el maní, así como en la producción y procesamiento de alimentos, se generen sustancias mutagénicas y cancerígenas (97, 98). Estas sustancias resultantes, que son formas activas de mutágenos, se clasifican como genotoxinas, porque se unen a la molécula del ADN, modifican sus bases y favorecen la aparición de mutaciones. Así, provocan la inactivación de genes supresores y la pérdida de la heterocigosidad (LOH, del inglés *loss of heterozygosity*) o la activación de protooncogenes que pueden promover la inestabilidad genómica, la aneuploidía y la transformación maligna seguida de la progresión neoplásica (99-101).

En cuanto a la actividad física, se la entiende como todo movimiento que realiza un sujeto usando los músculos esqueléticos y que permite el gasto de energía, como hacer ejercicio o practicar un deporte, realizar tareas domésticas, bailar, caminar, entre otras actividades (102). Existen estudios epidemiológicos que demuestran que, a través de diferentes mecanismos, la merma o la falta total

de actividad física contribuyen de manera estrecha con la formación de tumores. Es bueno recordar que su ausencia o escasez contribuyen al sobrepeso y que este actúa alterando y aumentando la actividad endocrina (estrógenos, andrógenos, insulina y factores de crecimiento), lo que, a su vez, permite el desarrollo de tumores (103).

Luego se encuentran las infecciones, como otro factor predisponente biológico para el cáncer. En el año 2017,[5] el Instituto Nacional del Cáncer de EE. UU. afirmó que existen numerosos agentes infecciosos que contribuyen a ciertos tipos de cáncer (104, 105). De esta manera, se han identificado agentes etiológicos infecciosos, tales como las bacterias, virus, hongos y parásitos, que están asociados a las causales de cáncer. Algunos de ellos son: a) *Helicobacter pylori* (cáncer gástrico), b) *Chlamydia pneumoniae* (cáncer de pulmón), c) *Chlamydia trachomatis* (cáncer de cuello de útero), d) virus de Epstein-Barr (EBV) (carcinoma nasofaríngeo, linfoma de Burkitt, cáncer gástrico, linfoma de Hodgkin y cánceres de lengua y amígdalas palatinas), e) *Salmonella typhi* (colangiocarcinoma), f) *Streptococcus bovis* (cáncer colorrectal), g) *Borrelia burgdorferi* (linfoma), h) *Human herpesvirus 8* (HHV-8) (o virus del sarcoma de Kaposi, está relacionado con el sida), i) virus del papiloma humano (VPH) (cáncer de cuello uterino, y cánceres de lengua, amígdalas y esófago), j) virus del papiloma humano T16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 y 59 (adenocarcinomas escamosos [90 %] y cervicales [75 %]), k) virus de las hepatitis B, C y D (VHB, VHC, VHD, la hepatocarcinoma primaria 80 % por VHB), l) virus del polioma de células de Merkel humano, retrovirus linfotrópico de células T humanas tipo 1 (HTLV-1) y el virus de la inmunodeficiencia humana (VIH) tipo 1 y 2 (leucemia, cáncer de

5 Fuente: https://tinyurl.com/4cbj89rj

pulmón y linfomas), m) *Schistosoma hematobium* (cáncer gástrico), n) *Trematodos hepáticos* (cáncer de vejiga), o) *Opisthorchis viverinni y Clonorchis sinensis* (colangiocarcinomas y carcinomas hepatocelulares).

En Argentina, las estimaciones realizadas en el año 2018 mostraron que cerca de diez mil casos de cáncer pueden ser atribuidos a estas infecciones (8,1 %), lo que las ubica en un lugar intermedio a nivel mundial. Con relación a Sudamérica, la IARC estimó ciento diez mil casos de cáncer atribuidos a infecciones (en el año 2018), los cuales se debieron principalmente a VPH (43,6 %) y a *Helicobacter pylori* (38,8 %), seguido del virus de hepatitis C (6,8%) y virus de hepatitis B (3,5%). En cuanto al sexo, el VPH explicó el 64 % de los cánceres atribuibles a infecciones en mujeres y el *Helicobacter pylori* explicó el 60,5 % de los cánceres atribuibles a infecciones en varones (106-108).

Por último, específicamente se hallan las hormonas utilizadas en los anticonceptivos orales. Actúan como mediadores químicos del organismo y son producidas por las glándulas endocrinas del cuerpo humano. Su transporte es a través del torrente sanguíneo a todos los componentes del cuerpo, ya sean células, tejidos, órganos (entre otros). Se definen como medicamentos del tipo hormonal, de administración por vía oral, con la finalidad de tener control sobre la natalidad, y se componen de versiones sintéticas de las hormonas femeninas naturales, que son el estrógeno y la progesterona (estrógeno y progestina).[6]

Respecto al vínculo entre el uso de anticonceptivos orales y el riesgo de desarrollar cánceres de mama, cuello de útero, endometrio, ovario, colon y recto se ha informado que efectivamente existe una asociación entre el uso de dichos anticonceptivos y los cánceres de mama y cuello

6 Fuente: https://tinyurl.com/777uw6ar

uterino a largo plazo. Asimismo, se ha demostrado que el uso de esta medicación reduce los riesgos de padecer cánceres de endometrio, ovario, colon y recto (109, 110).

2

Aprendizaje automático (*machine learning*)

La inteligencia artificial (IA) fue definida en 1956 por John McCarthy, Marvin Minsky y Claude Shannon como "la ciencia e ingenio de hacer máquinas inteligentes, especialmente, programas de cálculo inteligentes". En 1997, la IA fue presentada con *Deep Blue*, una computadora de IBM que fue capaz de ganarle en una partida de ajedrez al entonces campeón mundial Gari Kaspárov. Este hito marcó un antes y un después en el desarrollo de esta tecnología.[7]

El aprendizaje automático es un campo de la inteligencia artificial que puede, mediante una gran base de datos y algoritmos, aprender e identificar patrones de comportamiento para predecir posibilidades de ocurrencia casi sin intervención del humano. Puede realizar tareas sobre seguridad informática, traducción, reconocimiento del habla, tendencias y recomendaciones acerca de un producto o servicio.

El aprendizaje supervisado es una subcategoría del *machine learning*. Utiliza una serie de datos etiquetados para enseñar a los modelos a producir el resultado deseado. Estos datos contienen información sobre resultados correctos para que el modelo pueda aprender de la mejor manera.

---

7 https://tinyurl.com/yxsatk8n

El aprendizaje no supervisado utiliza datos sin etiquetar. A partir de esos datos, descubre patrones que pueden ayudar a resolver algún tipo de problema (111).

Los modelos se clasifican en a) aprendizaje supervisado, b) aprendizaje no supervisado, c) aprendizaje semisupervisado y d) aprendizaje reforzado. Es decir, estos algoritmos reciben y analizan los datos de entrada y predicen los valores de salida. En la medida que se introducen más datos en los algoritmos, en un principio las predicciones pueden ser más precisas, lo que da lugar al desarrollo de "inteligencia".

Los modelos mencionados se definen como:

a. **Aprendizaje supervisado**. Se da cuando el operador entrena al algoritmo y carga los datos de las entradas y salidas esperadas. Como el operador conoce las respuestas, corrige las predicciones realizadas por el algoritmo y es por ello que se alcanza una mayor precisión predictiva.
b. **Aprendizaje no supervisado.** A diferencia del anterior, el operador no da la instrucción al algoritmo ni tiene la respuesta. Por el contrario, el propio algoritmo establece los patrones mediante el análisis de los datos existentes. En este proceso se deja que el algoritmo de aprendizaje no supervisado interprete grupos en los cuales se puedan clasificar los datos sobre la base de los patrones encontrados.

Ejemplos de algoritmos de agrupamiento incluyen el algoritmo de k-means, el agrupamiento jerárquico y el agrupamiento espectral. Otra estrategia que hace uso del aprendizaje no supervisado es la reducción de dimensionalidad (*dimensionality reduction*), que se utiliza para reducir la cantidad de variables o características en un conjunto de datos, mientras se conserva la mayor parte de

la información relevante. Esto se logra mediante la extracción de características importantes o mediante la transformación de los datos en un espacio de menor dimensión. Las técnicas comunes de reducción de dimensionalidad incluyen el análisis de componentes principales (PCA).

c. **Aprendizaje semisupervisado (*semi–supervised learning*).** Se utiliza un conjunto de datos que contiene tanto ejemplos etiquetados como no etiquetados. La idea es aprovechar la información de los ejemplos etiquetados para guiar el aprendizaje de los ejemplos no etiquetados. En este enfoque, los algoritmos intentan aprender patrones y estructuras utilizando tanto la información explícita de las etiquetas como la información implícita en los datos no etiquetados. El objetivo es mejorar el rendimiento del modelo al aprovechar la cantidad, potencialmente mayor, de datos no etiquetados disponibles.
d. **Aprendizaje reforzado (*reinforcement learning*).** Es un enfoque por el cual un agente de aprendizaje interactúa con un entorno y aprende a tomar decisiones óptimas a través de su retroalimentación. En lugar de proporcionarse ejemplos etiquetados de antemano, el agente toma acciones en el entorno y recibe recompensas o penalizaciones según el resultado de sus acciones. El objetivo del aprendizaje por refuerzo es aprender una política o estrategia que maximice las recompensas a largo plazo. Es un enfoque comúnmente utilizado en problemas de toma de decisiones secuenciales, como los juegos, la robótica y el control de procesos (111, 112).

2.1. Algoritmos

Los algoritmos de aprendizaje automatizado que se emplean en este trabajo son:

- **Algoritmos de clasificación binaria**: representan una puntuación de predicción que indica la certeza del modelo usado. La tarea consiste en clasificar los datos como positivos y negativos. Para ello, el algoritmo contará con un umbral estipulado de corte para poder clasificar si un dato pertenece a la clase positiva (puntuación superior al umbral) o, en caso contrario, a la clase negativa (puntuación inferior al umbral). Por ejemplo, en la clasificación binaria solo se puede asignar un valor de 0 o 1. Un caso típico es la detección del email *spam*, donde: *spam* → será etiquetado con un 1; o no es *spam* → será etiquetado con un 0.
- **Clasificadores multiclase**: a diferencia de la categorización binaria, en la clasificación multiclase no hay un umbral (puntuación máxima o mínima) para asignar etiquetas o clases. Para llevar a cabo este proceso, este algoritmo calcula el promedio de la métrica entre todas las clases, lo que permite obtener una métrica de promedio macro o media ponderada. La métrica utilizada para evaluar su capacidad predictiva es la F1.
- ***Support vector machines* (SVM)**: es un algoritmo que puede utilizarse tanto con modelos de clasificadores como de regresión (estos últimos se explicarán en páginas posteriores). Mapea los datos de entrada a un espacio de mayor dimensión para encontrar un hiperplano lineal que maximice la

separación entre las clases. Además, puede emplear funciones de kernel para trabajar con patrones no lineales (112).

- **Algoritmos de árbol de decisión**: simulan la estructura de un árbol, donde cada bifurcación representa un resultado plausible de una decisión. Los nodos del árbol corresponden a pruebas realizadas sobre una variable específica, y las ramas indican los posibles resultados de dichas pruebas (113). Este algoritmo no fue utilizado para el análisis en este trabajo, pero sí lo fueron los algoritmos de *decision forest* y *random forest*, que pertenecen a este grupo. Estos se describen a continuación.
    - *Decision forest* es un algoritmo de aprendizaje supervisado que utiliza múltiples árboles de decisión para realizar predicciones. Cada árbol en el bosque es una comparación donde se entrena al algoritmo con una muestra aleatoria del conjunto de datos y toma decisiones basadas en características relevantes. Luego, las predicciones de todos los árboles se combinan para llegar a una predicción final. La fuerza de esta técnica radica en su capacidad para manejar características de entrada tanto numéricas como categóricas, así como para lidiar con el sobreajuste y la falta de generalización. Este algoritmo puede aplicarse sobre variables de tipo continuas o discretas y datos etiquetados o no etiquetados.
    - *Random forest* es un algoritmo dentro de los bosques de decisión que utiliza un enfoque de "ensamble", en el que se generan

múltiples árboles utilizando muestras aleatorias del conjunto de datos y subconjuntos aleatorios de características. Las predicciones de los árboles individuales se combinan mediante votación o promedio para obtener una predicción final (114, 115).

- **Algoritmos de reducción de dimensionalidad**: admiten la reducción de variables, trabajando con aquellas que permiten acceder a la información más exacta. Dentro de esta categoría se destacan los *análisis de correlación*, en los cuales la matriz calcula la relación entre las diversas variables y, a partir de esto, toma decisiones. El análisis de correlación es una técnica estadística utilizada para medir la relación entre variables. Su intensidad está determinada por el coeficiente de correlación, que varía entre -1 y +1. Es decir, se pueden presentar correlaciones negativas o positivas.

A continuación, se encuentran los *análisis de componentes principales* (*principal component analysis* o PCA), un método estadístico cuya utilidad radica en la reducción de la dimensionalidad de la base de datos (BDD) con la que se está trabajando. Se utiliza cuando es necesario simplificar la base de datos, por ejemplo, al elegir un número reducido de predictores para pronosticar una variable objetivo. Esta técnica forma parte de los algoritmos de aprendizaje no supervisado y solo emplea variables numéricas.

Por otra parte, están los *análisis de la ganancia de información* basada en la entropía de Shannon. Este análisis es particularmente útil cuando no es evidente cuánta información aporta cada atributo al resultado final. Es un buen método para obtener la relación

entre las variables independientes y la variable objetivo, utilizando conceptos de entropía de la termodinámica aplicados a la información (Shannon, 1948). En este sentido, se puede obtener un *ranking* ordenado de las variables según la magnitud de su aporte de información hacia la variable objetivo. Se utiliza como un método alternativo al cálculo de las correlaciones estadísticas.

La entropía mide el grado de desorden o incertidumbre en un conjunto de datos, y se define como "pi, la probabilidad relativa de aparición de la propiedad i en el conjunto de datos". Por ejemplo, p1 podría ser la probabilidad de tener cáncer de pulmón y p2 la probabilidad de no tenerlo en una población determinada. En la entropía de la teoría de la información vemos que se mide la confusión o el desorden de los datos que pertenecen a un conjunto para una variable aleatoria (VA) discreta con densidad de probabilidad p. Este valor indica la medida del desorden de la variable, que varía entre 0 y 1. Un valor de 0 indica un orden total, mientras que cuanto más cerca esté de 1, mayor será el desorden (116).

Fórmula 1. Fórmula de la entropía

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

Fuente: adaptado de T. M. Cover, J. A. Thomas. *Elements of Information Theory*. 2nd. ed. Wiley, 2006.

Entonces, p es la probabilidad relativa de aparición de la propiedad x en el conjunto de datos. El logaritmo en base 2 se refiere a la cantidad de información

codificada en bits (0 y 1). El valor calculado aquí es la medida del desorden de la variable explicada (la variable "objetivo" o "*target*"), con un valor que va de 0 a 1. Un valor de 0 indica un orden total, mientras que cuanto más cerca esté de 1, mayor será el desorden.[8]

Dentro de estos análisis de la ganancia de información basada en la entropía de Shannon encontramos los **algoritmos de regresión logística**, definidos como los métodos utilizados en el aprendizaje automático para modelar y predecir relaciones entre variables numéricas. Estos buscan encontrar la función matemática que se ajuste mejor a los datos observados, con el objetivo de poder realizar predicciones. Las técnicas pueden ser lineales o no lineales, y se basan en diversos algoritmos, como la regresión lineal, polinomial, logística, regresión de vecinos más cercanos, entre otros. Son ampliamente utilizadas en diversas áreas (economía, ingeniería y medicina) para comprender relaciones causales, hacer pronósticos y tomar decisiones informadas basadas en datos numéricos.

En particular, la regresión logística es un algoritmo de clasificación que utiliza una función logística para modelar la relación entre variables predictoras y la probabilidad de pertenecer a una clase específica. Los coeficientes se ajustan mediante optimización de máxima verosimilitud y se establece un umbral para clasificar nuevas instancias. Es ampliamente utilizado en problemas de clasificación binaria debido a su simplicidad, interpretabilidad y buen rendimiento (116).

---

8 T. M. Cover, J. A. Thomas (2006). *Elements of Information Theory*. 2nd. ed., p. 5. Wiley.

Por otro lado, se encuentra el **algoritmo bayesiano**, basado en el teorema de Bayes, utilizado para calcular las probabilidades de una partición cuando se tiene conocimiento de un evento específico que ya ha ocurrido. Este enfoque clasifica cada valor de manera independiente de los demás, y así proporciona una predicción basada en la probabilidad de que un dato pertenezca a una categoría específica, según el conjunto de datos inicial. Utiliza el teorema de Bayes para modelar y realizar inferencias probabilísticas, actualizando creencias previas mediante la evidencia observada. De esta manera, combina información previa con datos recién recibidos para generar estimaciones más precisas. Los algoritmos bayesianos emplean distribuciones de probabilidad para representar la incertidumbre y calculan la distribución *a posteriori* a partir de la evidencia y el modelo. Son especialmente útiles cuando los datos son escasos o la incertidumbre es alta, lo que los hace aplicables en diversas áreas, como la clasificación, regresión, agrupamiento y toma de decisiones bajo condiciones de gran incertidumbre (113).

Algoritmos de agrupación (clúster): su funcionamiento se basa en la búsqueda de grupos dentro de los datos de manera iterativa, asignando cada punto a uno de los grupos representados por la variable definida (variable K). Dentro de esta agrupación, existe un subalgoritmo denominado A priori, el cual es un algoritmo de asociación que utiliza *machine learning* para analizar datos y buscar patrones o coocurrencias en una base de datos. Es decir, con base en el análisis de los datos obtenidos, evalúa la probabilidad de que otro dato sea seleccionado. Este tipo de algoritmo es ampliamente utilizado, por ejemplo, en el ámbito comercial, para predecir las compras de los clientes. Esto significa

que, si un cliente ha comprado el artículo 1, es probable que también compre el artículo 2. Las reglas de asociación se componen de dos partes:

1. Un antecedente (si): un elemento que se encuentra en el conjunto de datos.
2. Un consecuente (entonces): un elemento que resulta de la combinación con un antecedente (112).

Cabe aclarar que este algoritmo no se utilizó en el análisis del presente trabajo.

Además están los **algoritmos de redes neuronales**, que se comprenden como unidades organizadas en capas, donde cada una se conecta a las capas adyacentes. En estas redes participan un gran número de elementos interconectados que trabajan de manera conjunta para resolver situaciones específicas. Su funcionamiento se basa en el aprendizaje a partir de ejemplos y experiencia, lo que permite modelar relaciones no lineales en los datos, especialmente cuando la vinculación entre variables es compleja o difícil de comprender. (Ejemplo: perceptrón multicapa o *feedforward neural network*) (117).

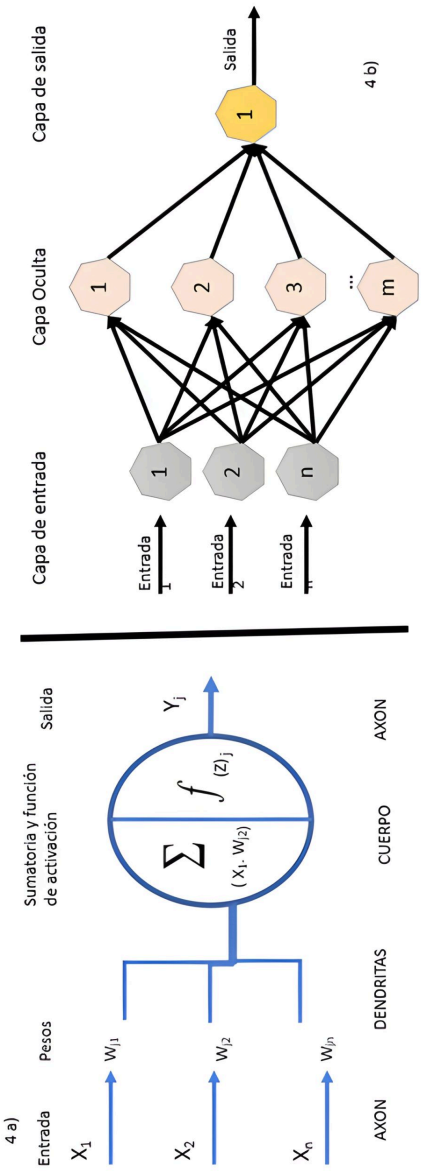


Figura 3. Esquemas de redes neuronales, basados en *An Introduction to Neural Networks*, **C. C. Aggarwa**

En el aprendizaje supervisado, la regulación o selección de un algoritmo para obtener mejores resultados puede ser el origen de un sesgo, debido a la participación humana. Por otro lado, en el aprendizaje no supervisado, las posibilidades de sesgos son menores porque el operador solo introduce los datos, y así permite que la máquina seleccione entre los algoritmos aprendidos y decida cuál es el más adecuado para analizar los datos y variables.

Existen métricas para la evaluación de los algoritmos, los cuales son medidores que permiten caracterizar su funcionamiento. Entre estas métricas se incluyen:

a. Especificidad (*accuracy*): mide la proporción de predicciones correctas con respecto al número total de predicciones. Además, refleja los casos negativos clasificados correctamente por el algoritmo.
b. Matriz de confusión: es una tabla que registra y visualiza las predicciones sobre un conjunto de datos categorizados, evaluando el rendimiento del modelo en términos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Es decir, cada columna representa el número de predicciones realizadas para cada clase, mientras que cada fila indica las instancias en la clase real. Esta métrica muestra el nivel de error y acierto del modelo utilizado de manera numérica.

Figura 4. Representación de la matriz de confusión

Matriz de confusión

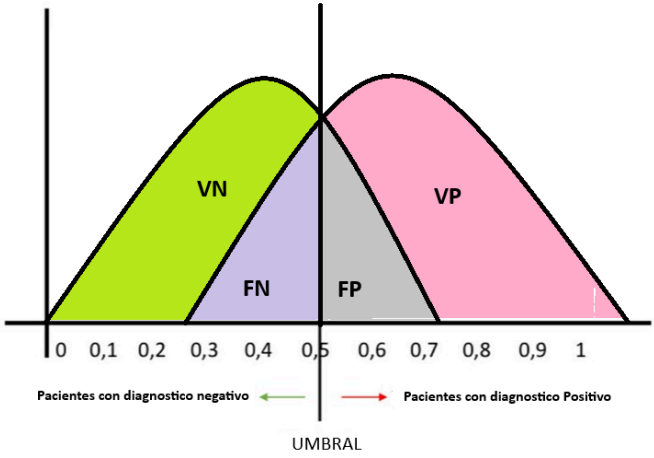
		Predicción	
		Positivo	Negativo
Actual			
Verdadero		Positivo Verdadero (TP)	Falso Negativo (FN)
Falso		Falso Positivo (FP)	Negativo Verdadero (TN)

Fuente: Elaboración propia.

c. El área bajo la curva (AUC): es una métrica que mide la capacidad del modelo para distinguir entre clases positivas y negativas. Cuanto mayor sea el valor del AUC, mejor será el rendimiento del modelo. Su interpretación se basa en los siguientes casos:

- AUC = 1: indica una clasificación perfecta, en la que todas las clases positivas y negativas son correctamente identificadas.
- AUC = 0: representa un escenario en el que todas las predicciones son erróneas, se clasificaron todos los positivos como negativos y viceversa.
- 0,5 < AUC < 1: el modelo tiene cierta capacidad para discriminar entre las clases positivas y negativas, logra resultados efectivos en término de verdaderos positivos (TP) y verdaderos negativos (TN).
- AUC = 0,5: el modelo no tiene capacidad de distinción entre clases, y su predicción es equivalente a una decisión aleatoria (118).

Figura 5. Representación del área bajo la curva



d. Precisión y sensibilidad *(recall)*: permiten evaluar la calidad de un modelo en tareas de clasificación. La precisión mide la proporción de predicciones relevantes entre las recuperadas, mientras que la sensibilidad (o *recall*) muestra la proporción de casos positivos que fueron correctamente identificados por el algoritmo.

Fórmula 2. Fórmula para sensibilidad (*recall*) = positivos reales (TP: *true positives*) sobre positivos reales (TP) más negativos falsos (FN: *false negatives*)

$$RECALL = \frac{TP}{TP + FN}$$

Fuente: elaboración propia.

e. Puntuación F1 (F1 *score*): es una métrica de evaluación en el aprendizaje automático que representa el promedio armónico entre la precisión y la sensibilidad

(*recall*). Esta puntuación se obtiene calculando la precisión y el *recall* (sensibilidad) del modelo, y refleja cuántas veces realizó predicciones correctas en relación con el conjunto de datos analizado (117, 119).

Fórmula 3

$$Puntuación\ F1 = \frac{2.Precisión.Recall}{TP + FN}$$

Fuente: elaboración propia.

3

Materiales y métodos

El presente estudio adoptó un diseño metodológico observacional, de cohorte, mixto y epidemiológico, utilizando una muestra no probabilística aleatoria. La muestra estuvo conformada por 2.000 habitantes de la Ciudad Autónoma de Buenos Aires (CABA) y del primer cordón del Gran Buenos Aires (GBA), con edades de entre 18 y 75 años, garantizando una representación equitativa del 50 % por sexo en los grupos de casos y controles. Las unidades de análisis incluyeron a 333 sujetos con diagnóstico confirmado de cáncer mediante histología, que formaron el grupo "CASO", y a 1.276 sujetos seleccionados aleatoriamente de CABA y el primer cordón del GBA, que conformaron el grupo "CONTROL".[9]

Asimismo, como criterios de inclusión, se estableció que los participantes fueran habitantes de CABA y del primer anillo del GBA, con edades entre 18 y 75 años, que aceptaron responder la encuesta y firmaron el consentimiento informado. También se incluyeron aquellos habitantes dentro del mismo rango etario que tuvieran un diagnóstico positivo de cáncer por histología ("caso"), así como a personas sanas en apariencia al momento de la toma de la encuesta ("control").

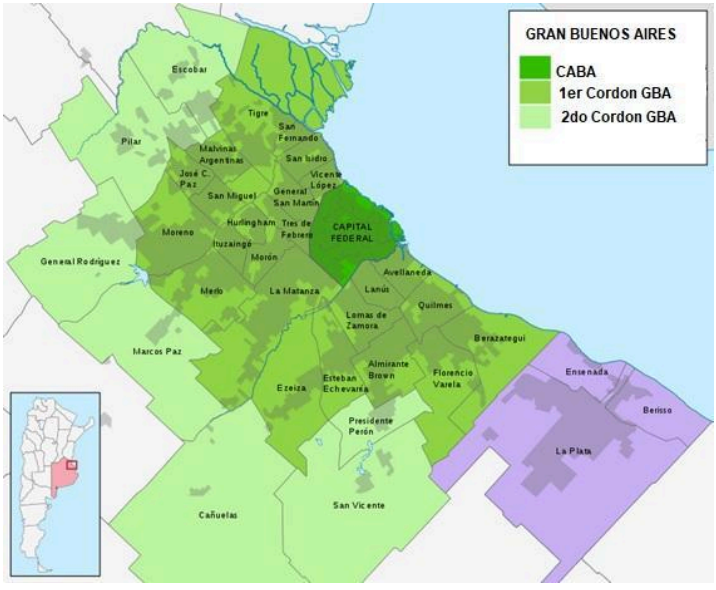
---

9 La denominación "caso" y "control" solo es para identificar la diferencia en la muestra obtenida, no hace referencia a la denominación metodológica de los estudios analíticos (casos y controles).

En cambio, fueron excluidos los habitantes de CABA y del primer anillo del GBA, con edades de entre 18 y 75 años, que no hayan firmado el consentimiento informado, así como aquellos que no aceptaron responder la encuesta; y, por último, se prescindió de las personas que fueran personal de salud con título intermedio (técnico), superior o universitario.

Este trabajo se realizó en la Ciudad Autónoma de Buenos Aires y el primer cordón que la rodea. El primer cordón abarca los partidos de Avellaneda, Lanús, Lomas de Zamora, La Matanza (en parte), Morón, Tres de Febrero, San Martín, Vicente López y San Isidro.

Figura 6



Fuente: Adaptado del INDEC. Última actualización: 31/1/2023.

Se utilizó como instrumento de recolección de datos una encuesta anónima, mixta y de opciones múltiples, a partir de la cual se obtuvieron 2.600 respuestas. Para acceder a la población definida como grupo "caso", se solicitó autorización a los comités de ética de los hospitales participantes, incluyendo el Hospital de Clínicas, el Hospital María Curie, el Centro Médico Austral, el Hospital Narciso López de Lanús y el Hospital Luisa C. de Gandulfo de Lomas de Zamora. Los casos correspondieron a pacientes tratados en estos hospitales durante el período del estudio, que abarcó los años 2016 a 2018.

La población del grupo "control" estuvo compuesta por personas de las mismas comunidades donde se originaron los casos, pertenecientes al mismo ámbito geográfico y temporal del estudio. Los controles fueron seleccionados sin considerar sus hábitos o antecedentes, con el objetivo de minimizar posibles sesgos muestrales.

3.1. Análisis de los resultados[10]

En el presente trabajo de investigación, los resultados se obtuvieron mediante el análisis de un conjunto de datos procesados utilizando técnicas de aprendizaje automático (*machine learning*) dentro del campo de la inteligencia artificial.[11]

Se aplicó una encuesta y se realizó la codificación de los tipos de cáncer y enfermedades utilizando la Clasificación Internacional de Enfermedades (CIE-10) (véase Tabla AI2 en Anexo I). También se registraron antecedentes

---

10 Véase Anexo I para la codificación de las variables.

11 Aclaración: la base de datos digital fundada en la encuesta estuvo a cargo del Ing. Claudio Milio, quien luego trabajó con el conjunto de datos junto con el Lic. Jorge Kamlosfky.

familiares y el uso de medicamentos (véase Listado N.° 1 en Anexo I). Una vez creada y completada la base de datos, se procedió a su depuración aplicando los criterios de inclusión y exclusión previamente definidos. Se eliminaron los registros con anomalías o datos incompletos. En el caso de un registro sin información de edad, pero con diagnóstico confirmado de cáncer, se asignó la edad promedio del conjunto de datos, que fue de 40 años.

- Cánceres (variables *target* u objetivo: casos)

C-18: colon, C-34: pulmón, C-44: piel, C-50: mama, C-55: cuello de útero, C-61: próstata, C-91: linfoma, C-95: leucemia.

- Cánceres en general (todos de los que se obtuvieron casos en la muestra)

C-06 : boca y lengua, C-18: colon, C-55: cuello de útero, C-63: epididimal (testículo), C-16: estómago, C-22: hígado, C-39: laringe, C-91: linfoma, C-95: leucemia, C-81: linfoma no Hodgkin, C-76: liposarcoma, C-50: mama, C-38: mediastino, C-43: melanoma, C-45: mesotelioma, C-41: osteosarcoma (columna vertebral), C-56: ovario, C-25: páncreas, C-44: piel, C-61: próstata, C-34: pulmón, C-20: recto, C-68: riñón, C-46: sarcoma de Kaposi; C-62: testículo, C-55: útero, C-67 vejiga, y C-23: vesícula.

Tabla 1. Cantidad de cánceres por diagnóstico

Suma diag. C06=4	Suma diag. C16=7	Suma diag. C18=43	Suma diag. C20=2
Suma diag. C22=2	Suma diag. C23=1	Suma diag. C25=4	Suma diag. C34=21
Suma diag. C38=6	Suma diag. C39=2	Suma diag. C41=1	Suma diag. C43=4
Suma diag. C44=19	Suma diag. C45=2	Suma diag. C46=2	Suma diag. C50=101
Suma diag. C55=30	Suma diag. C56=3	Suma diag. C57=1	Suma diag. C61=41
Suma diag. C62=3	Suma diag. C63=1	Suma diag. C71=1	Suma diag. C76=1
Suma diag. C81=1	Suma diag. C91=10	Suma diag. C95=14	

Fuente: elaboración propia: *output* pc.

- Cantidad de sujetos por cada análisis

Tabla 2. Cantidad de sujetos dentro del grupo "casos" (suma diag. cáncer) y grupo "control" (suma diag. no cáncer) para el primer análisis de resultados

Suma diag. no cáncer = 1276	Suma diag. cáncer = 329

Fuente: elaboración propia: *output* pc.

3.1.1. Primer análisis de los resultados

Se identificaron las variables del trabajo y se obtuvo un total de noventa y una. Se aplicó un programa desarrollado en Python que utiliza conceptos de la teoría de la información de Shannon, a partir del cual se realizó una selección de variables objetivo.

Con el apoyo de la herramienta *Azure Machine Learning*©, se generó un *ranking* de las variables independientes que mostraron una mayor correlación, tanto positiva como negativa, con la variable dependiente *cáncer* (también denominada "objetivo" o "*target*").

Para el análisis, se consideraron los 329 diagnósticos de cáncer registrados, de los cuales se seleccionaron ocho grupos con un número significativo de casos (véase Listado 1 de variables en el Anexo II) y un total de 1276 controles.

3.1.2. Segundo análisis de los resultados

El siguiente análisis se realizó para identificar posibles variables omitidas en el primer análisis, repitiendo el proceso previo de reconocimiento de variables, con un total de 849. De este modo, se pudo constatar que el conjunto de variables predictoras aumentó de 88 a 849 (véase Listado 2 de variables en el Anexo II).

Tabla 3. Cantidad de sujetos dentro del grupo "casos" (suma diag. cáncer) y grupo "control" (suma diag. no cáncer) para el segundo análisis de resultados

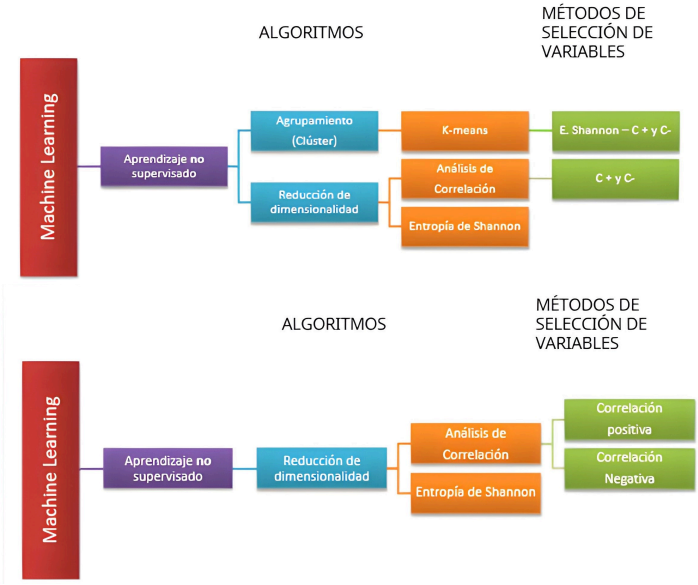
Suma diag. no cáncer = 1609	Suma diag. cáncer = 340

Fuente: elaboración propia: *output* pc.

4

Resultados de aprendizaje no supervisado

Figura 7. Primer y segundo análisis de resultados



Fuente: elaboración propia.

Primer análisis

4.1. Reducción de dimensionalidad

En el marco de este estudio, se aplicaron diversas técnicas de aprendizaje no supervisado para analizar y seleccionar las variables más relevantes. El primer paso consistió en realizar un análisis de reducción de dimensionalidad, con el fin de simplificar el conjunto de datos sin perder información esencial. Dentro de este proceso, se emplearon dos enfoques principales:

1. **Selección de variables por entropía de Shannon**, que permitió identificar aquellas variables más informativas en función de su grado de incertidumbre.
2. **Análisis de correlación**, donde se realizaron dos tipos de análisis:
    - Correlación positiva, para identificar las variables que mostraban una relación directa.
    - Correlación negativa, que permitió detectar variables con relaciones inversas.

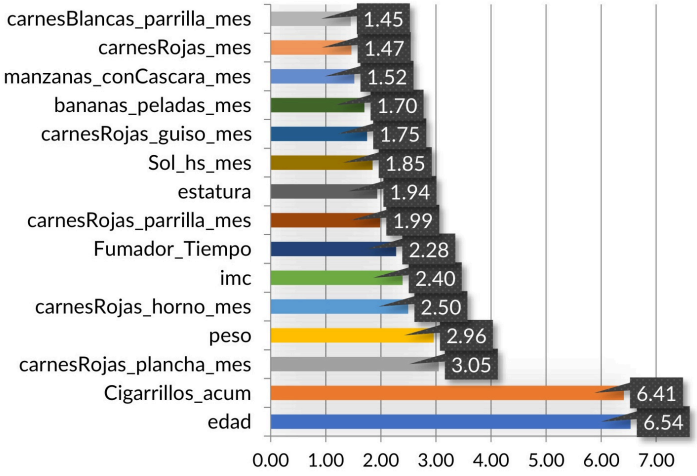
Posteriormente, se utilizó el algoritmo de clustering (k-means) para agrupar las variables y observar patrones o similitudes entre ellas. En esta fase, se repitieron tanto la selección de variables por entropía de Shannon como el análisis de correlación. A continuación, se presentan los resultados detallados de estas selecciones y análisis.

4.1.1. Variables seleccionadas por entropía de Shannon (teoría de la información)

Se presenta el *ranking* de las quince variables que más información aportan a las variables objetivo.

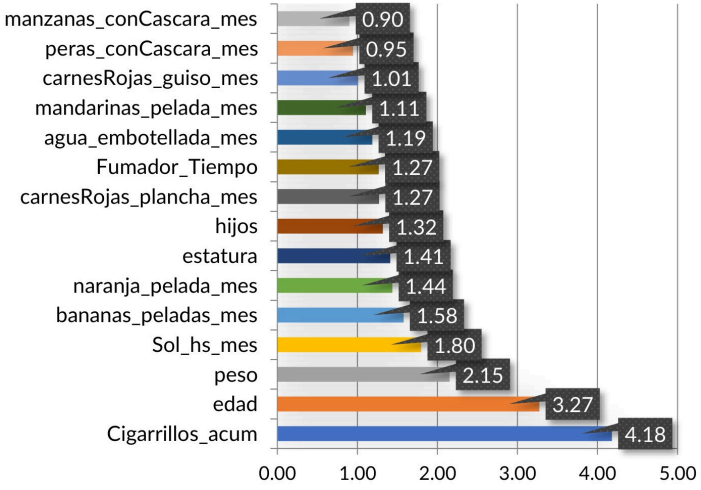
Descripción de las variables por diagnóstico

Figura 8. Gráfico de distribución de frecuencia según la entropía del cáncer de colon



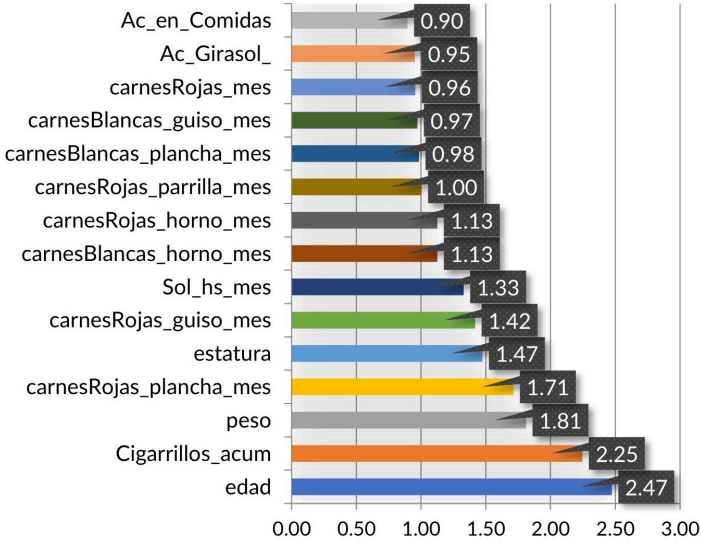
Para el cáncer de colon las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *tiempo de fumador*, *exposición al sol en horas por mes.* En las variables de hábitos alimentarios: *carnes rojas cocidas*, *plancha, horno, parrilla, guiso al mes*, *carnes blancas cocidas a la parrilla al mes* y *consumo de bananas, naranjas, mandarinas, peras y manzanas al mes*.

Figura 9. Gráfico de distribución de frecuencia según la entropía del cáncer de pulmón



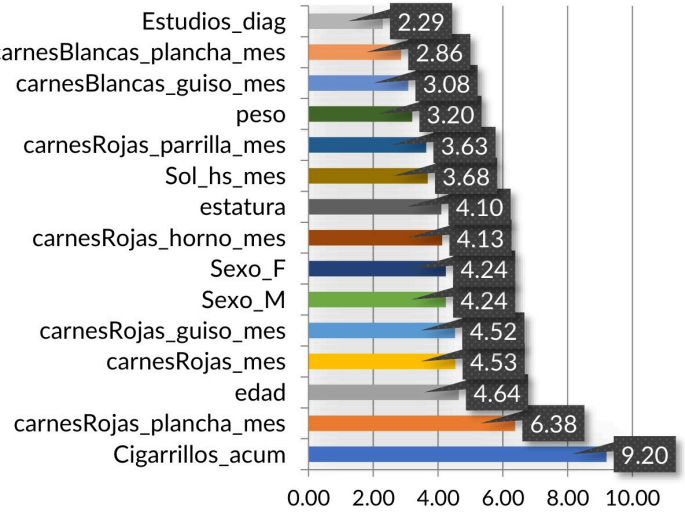
Para el cáncer de pulmón las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *tiempo de fumador*, *exposición al sol en horas por mes* e *hijos*. En las variables de hábitos alimentarios: *consumo de agua embotellada al mes*, *carnes rojas cocidas en plancha*, *guiso al mes*, *carnes blancas cocidas a la parrilla al mes* y *consumo de bananas, naranjas, mandarinas, peras y manzanas al mes*.

Figura 10. Gráfico de distribución de frecuencia según la entropía del cáncer de piel



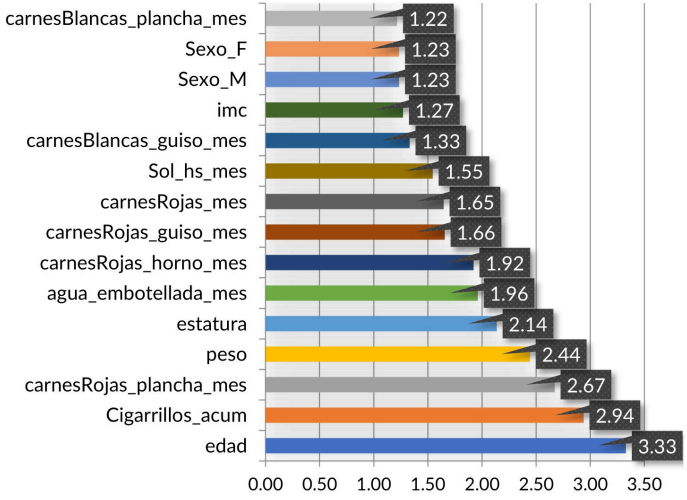
Para el cáncer de piel las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos, exposición al sol en horas por mes*. En las variables de hábitos alimentarios: *consumo de carnes rojas y blancas cocidas en plancha, guiso, horno, parrilla al mes*, y *consumo de aceite de girasol en las comidas.*

Figura 11. Gráfico de distribución de frecuencia según la entropía del cáncer de mama



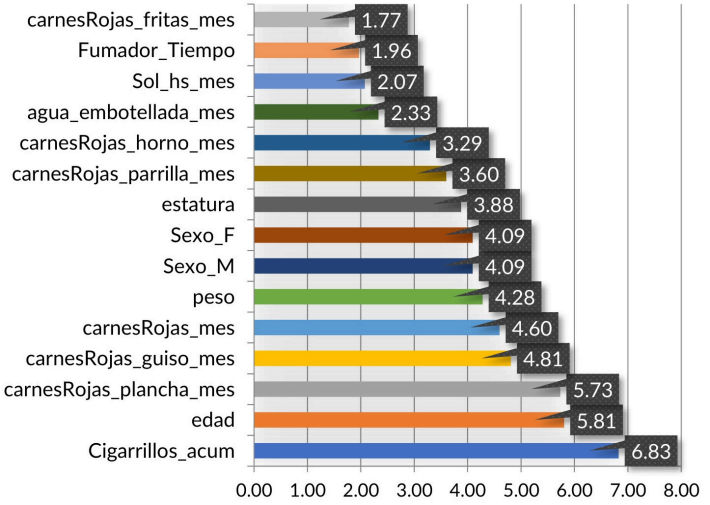
Para el cáncer de mama las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *exposición al sol en horas por mes* y en menor relevancia *realización de estudios diagnósticos*. En las variables de hábitos alimentarios: *consumo de carnes rojas cocidas en plancha, guiso, horno y parrilla al mes, consumo de carnes blancas cocidas en guiso y a la plancha al mes.*

Figura 12. Gráfico de distribución de frecuencia según la entropía del cáncer de cuello de útero



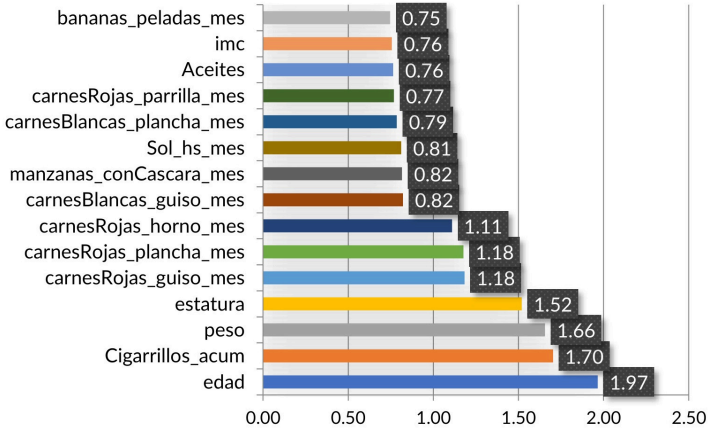
Para el cáncer de cuello de útero las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos, exposición al sol en horas por mes.* En las variables de hábitos alimentarios: *consumo de agua embotellada al mes, carnes rojas cocidas en plancha, horno y guiso al mes, carnes blancas cocidas en guiso y a la plancha al mes.*

Figura 13. Gráfico de distribución de frecuencia según la entropía del cáncer de próstata



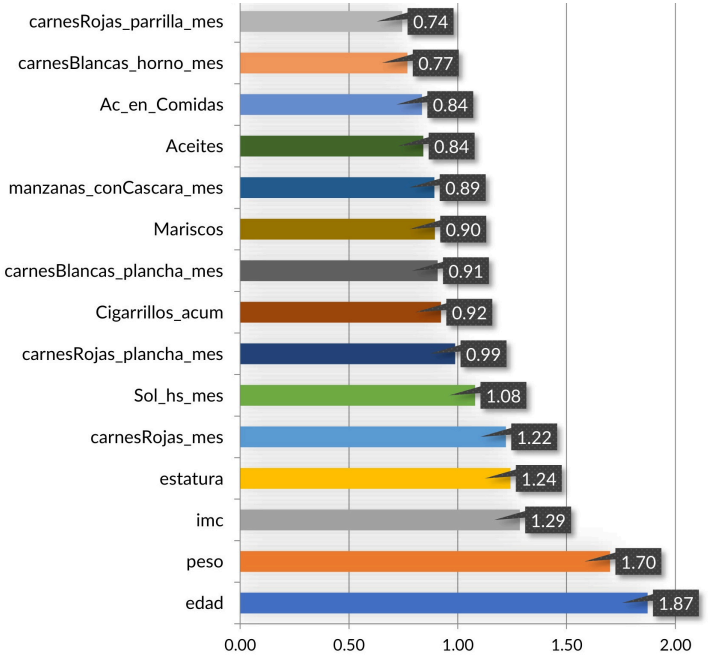
Para el cáncer de próstata las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *exposición al sol en horas por mes* y en menor relevancia *tiempo de fumador.* En las variables de hábitos alimentarios: *consumo de agua embotellada al mes* y *consumo de carnes rojas cocidas en plancha, guiso, parrilla, horno y fritas al mes*, *carnes blancas cocidas en guiso y a la plancha al mes*.

Figura 14. Gráfico de distribución de frecuencia según la entropía del cáncer linfoma



Para el linfoma las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *exposición al sol en horas por mes*. En las variables de hábitos alimentarios: *consumo de aceites* y *carnes rojas cocidas en guiso, plancha, horno y parrilla al mes*, *carnes blancas cocidas en guiso, plancha y a la parrilla al mes* y *consumo de manzanas y bananas al mes.*

Figura 15. Gráfico de distribución de frecuencia según la entropía de leucemia



Para leucemia las variables con mayor relación en entropía son: *exposición al sol en horas por mes* y *consumo acumulado de cigarrillos*. En las variables de hábitos alimentarios: *consumo de aceites* y *aceites en las comidas*, *mariscos* y *carnes rojas cocidas en la plancha* y *parrilla al mes*, *carnes blancas cocidas en guiso, plancha y al horno al mes* y *consumo de manzanas al mes.*

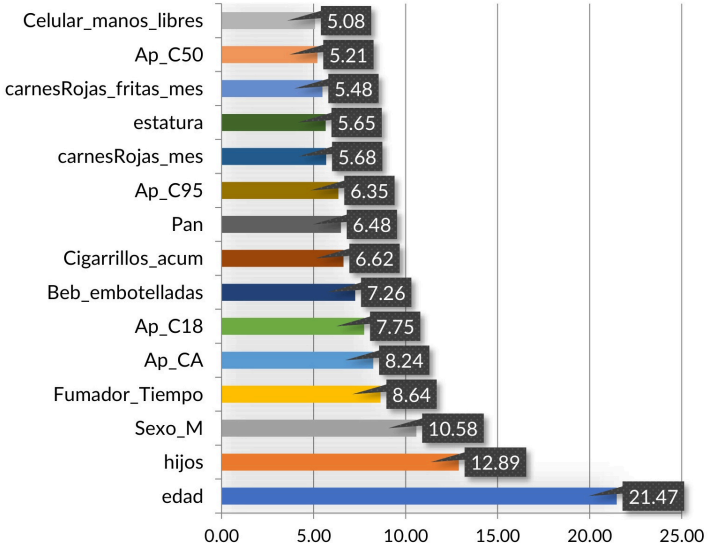
4.1.2. Variables seleccionadas por análisis de correlación

- **Correlación positiva** (se usa la correlación estadística)

Se presentan las quince variables con la mejor correlación positiva respecto al diagnóstico positivo de cada tipo de cáncer.

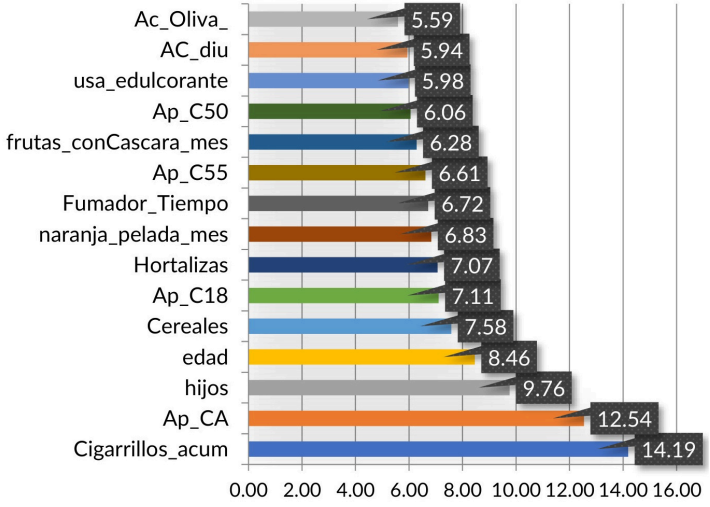
Descripción de las variables por diagnóstico

Figura 16. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de colon



La correlación positiva entre las variables en estudio con la variable objetivo muestra las de mayor importancia, las cuales son: *edad, número de hijos, sexo masculino* y *tiempo como fumador, consumo de bebidas embotelladas, consumo acumulado de cigarrillos y de pan*. Entre las variables finales del *ranking* están: *antecedentes personales de leucemia, consumo de carnes rojas fritas al mes, antecedente personal de cáncer de mama* y *uso de manos libres con el celular.*

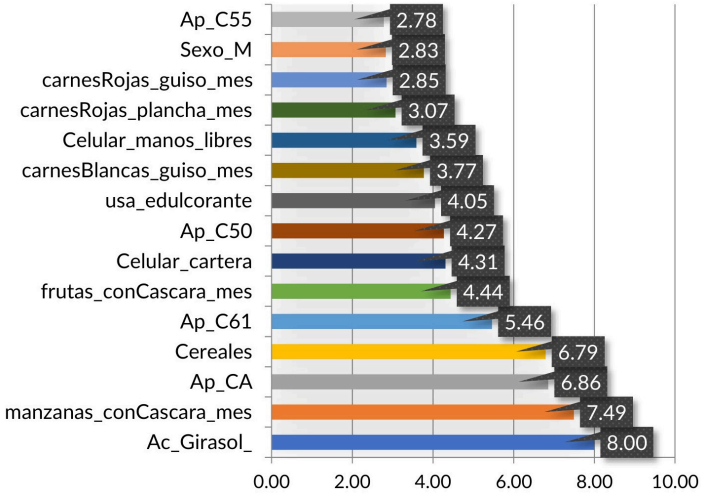
Figura 17. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de pulmón



La correlación positiva entre las variables con el cáncer de pulmón muestra que las de mayor importancia son: *consumo acumulado de cigarrillos, número de hijos, edad, consumo de cereales y hortalizas, antecedentes personales de cáncer de colon, cuello de útero* y *mama, tiempo como fumador, consumo de frutas con cáscara y naranjas al mes.*

Entre las variables finales del *ranking* están: *consumo de edulcorantes* y *aceite de oliva.* También está *uso del DIU como método anticonceptivo.*

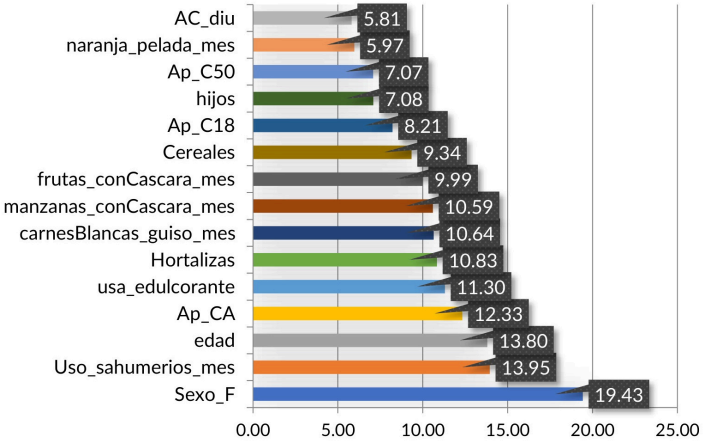
Figura 18. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de piel



La correlación positiva entre las variables con el cáncer de piel muestra que las de mayor importancia son: *consumo de aceite de girasol, cereales, frutas con cáscara* y *manzanas al mes, antecedente personal de cáncer de próstata y mama, portar el celular en la cartera* y *consumo de edulcorante.*

Entre las variables finales del *ranking* están: *consumo de carnes blancas y rojas en guiso, carnes rojas a la plancha, uso de las manos libres con el celular, sexo masculino* y *antecedente personal del cáncer de cuello de útero.*

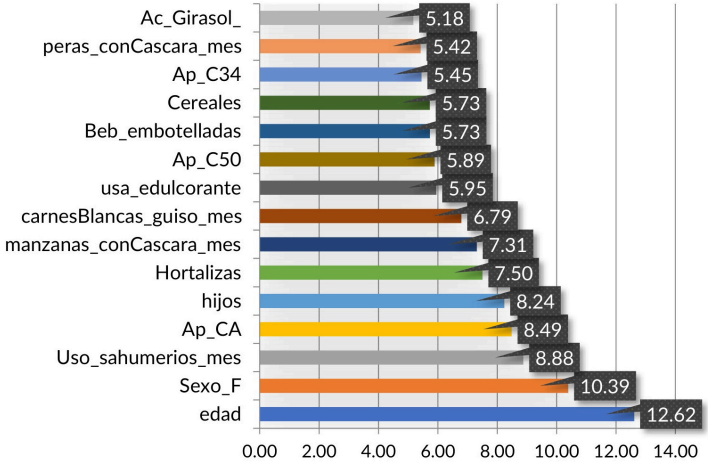
Figura 19. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de colon



La correlación positiva entre las variables con el cáncer de mama muestra que las de mayor importancia son: *sexo femenino, edad, número de hijos, antecedente personal de cáncer de colon, uso de sahumerios al mes, uso de edulcorantes, consumo de hortalizas, cereales, consumo de carnes blancas en guiso* y de *frutas y manzanas con cáscara al mes.*

Entre las variables finales del *ranking* están: *consumo de naranjas al mes* y *uso del DIU como anticonceptivo.*

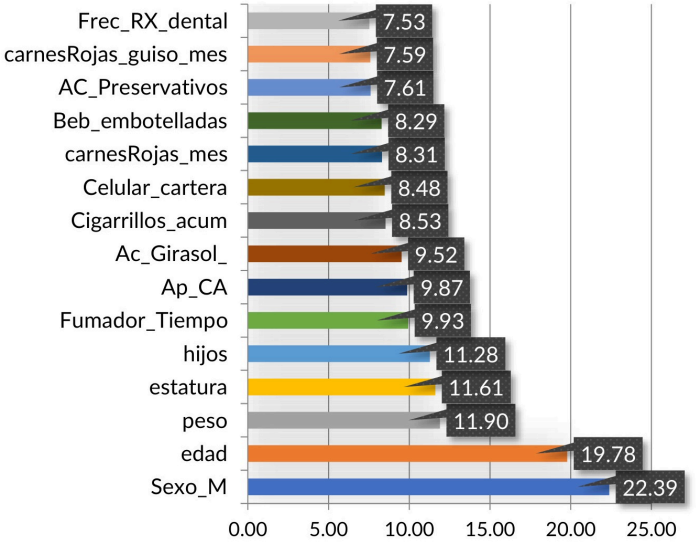
Figura 20. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de cuello de útero



La correlación positiva entre las variables con el cáncer de cuello de útero muestra que las de mayor importancia son: *edad, sexo femenino, uso de sahumerios al mes, número de hijos, uso de edulcorantes, consumo de hortalizas, cereales, consumo de carnes blancas en guiso* y de *frutas y manzanas con cáscara al mes, antecedente personal de cáncer de mama, consumo de bebidas embotelladas.*

Entre las variables finales del *ranking* están: *antecedente personal de cáncer de pulmón, consumo de peras al mes* y *uso de aceite de girasol.*

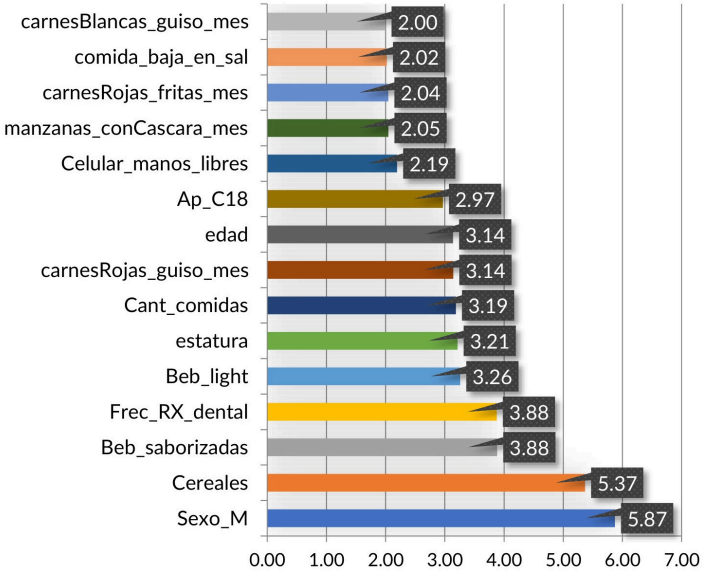
Figura 21. Gráfico de distribución de frecuencia según la correlación positiva del cáncer de colon



La correlación positiva entre las variables con el cáncer de próstata muestra que las de mayor importancia son: *edad, peso, estatura, número de hijos, tiempo de fumador, consumo de aceite de girasol, consumo acumulado de cigarrillos, portar el celular en la cartera, consumo de carnes rojas al mes* y *bebidas embotelladas.*

Entre las variables finales del *ranking* están: *uso del preservativo, consumo de carnes rojas en guiso al mes* y *realización de radiografías dentales.*

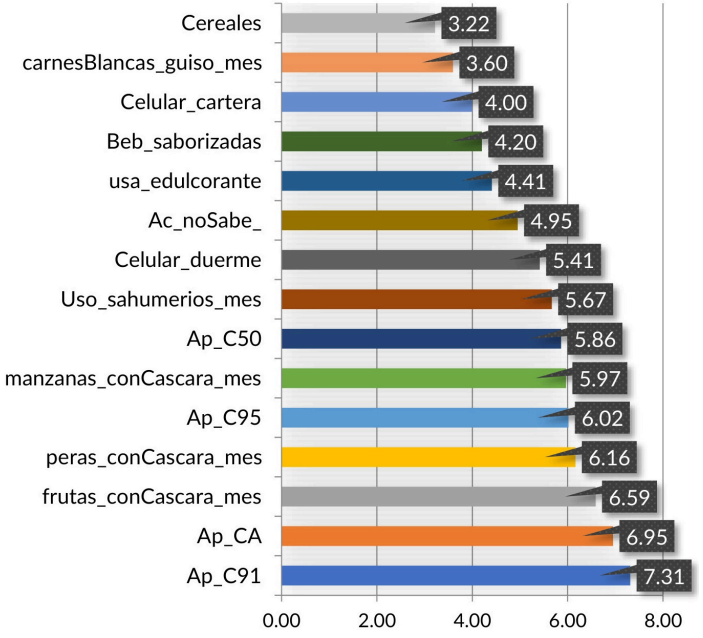
Figura 22. Gráfico de distribución de frecuencia según la correlación positiva del cáncer linfoma



La correlación positiva entre las variables con el linfoma muestra que las de mayor importancia son: *sexo masculino, consumo de cereales, bebidas saborizadas y light, realización de radiografías dentales, estatura, cantidad de comidas al día, consumo de carnes rojas en guiso por mes* y *edad.*

Entre las variables finales del *ranking* están: *antecedente personal del cáncer de colon, uso de manos libres con el celular, consumo de manzanas con cáscara al mes, comidas bajas en sal, carnes rojas fritas* y *carnes blancas en guiso al mes.*

Figura 23. Gráfico de distribución de frecuencia según la correlación positiva de leucemia



La correlación positiva entre las variables con la leucemia muestra que las de mayor importancia son: *antecedente personal de linfoma y cáncer de mama, consumo de frutas, peras y manzanas con cáscara al mes, uso de sahumerios al mes* y *dormir con el celular.*

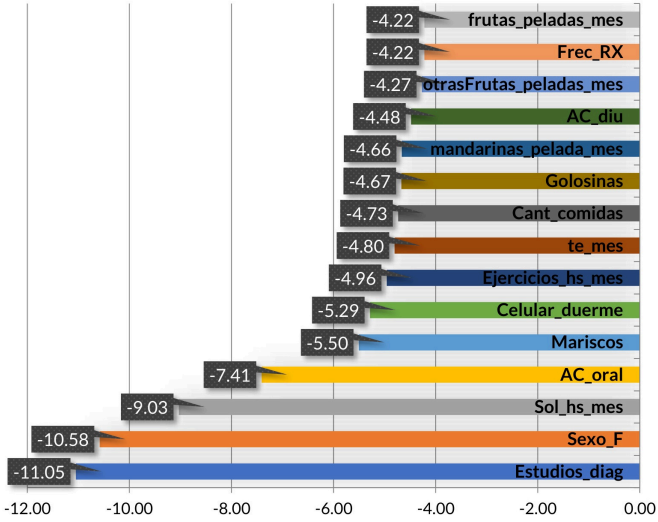
Entre las variables finales del *ranking* están: *desconocimiento sobre el uso de anticonceptivos, consumo de edulcorantes, bebidas saborizadas, cereales* y *consumo de carnes blancas en guiso por mes, portar el celular en la cartera.*

- **Correlación negativa**

Se presentan las quince variables que muestran la mejor correlación negativa con el diagnóstico de cada tipo de cáncer.

Descripción de las variables por diagnóstico

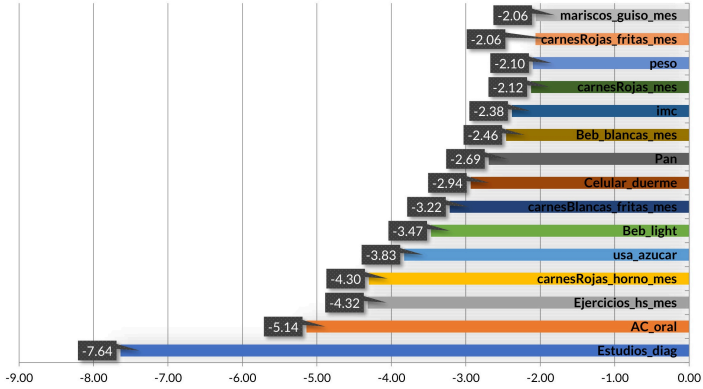
Figura 24. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de colon



La correlación negativa entre las variables con el cáncer de colon muestra que las de mayor importancia son: *estudios diagnósticos, sexo femenino, exposición a la luz solar al mes, uso de anticonceptivos orales, consumo de mariscos, dormir con el celular* y *ejercicio físico al mes.*

Entre las variables finales del *ranking* están: *consumo de té al mes, golosinas, frutas peladas y otras frutas al mes, uso de anticonceptivo DIU* y *frecuencia de radiografías.*

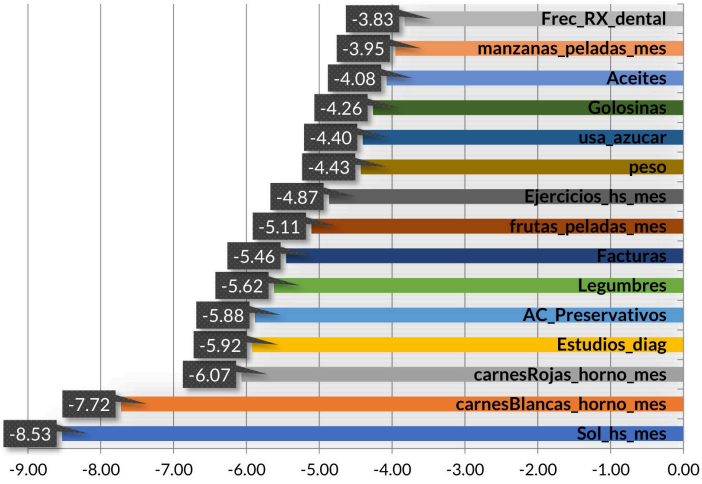
Figura 25. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de pulmón



La correlación negativa entre las variables con el cáncer de pulmón muestra que las de mayor importancia son: *estudios diagnósticos, uso de anticonceptivos orales, ejercicio físico al mes, consumo de carnes rojas al horno* y *carnes blancas fritas al mes, consumo de azúcar y bebidas light* y *dormir con el celular.*

Entre las variables finales del *ranking* están: *consumo de pan, bebidas blancas al mes, carnes rojas fritas* y *mariscos en guiso al mes, IMC y peso*.

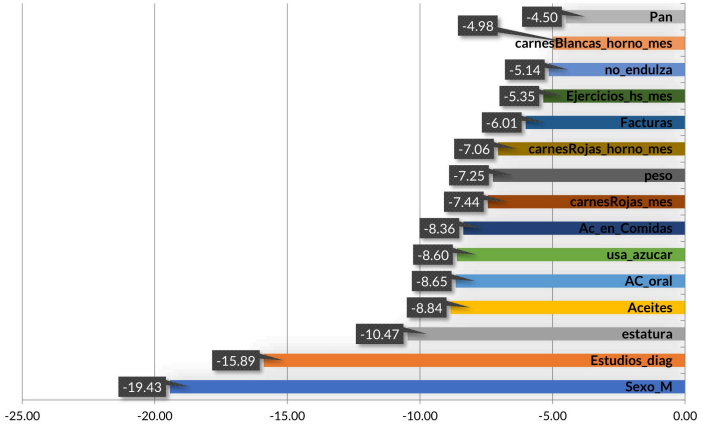
Figura 26. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de piel



La correlación negativa entre las variables con el cáncer de piel muestra que las de mayor importancia son: *exposición a la luz solar al mes, consumo de carnes blancas y rojas al horno por mes, estudios diagnósticos, uso de preservativos, consumo de legumbres, facturas, frutas peladas al mes* y *ejercicio físico al mes.*

Entre las variables finales del *ranking* están: *peso (en kg del sujeto), uso de azúcar, golosinas, aceites, manzanas peladas al mes* y *frecuencia de radiografías dentales.*

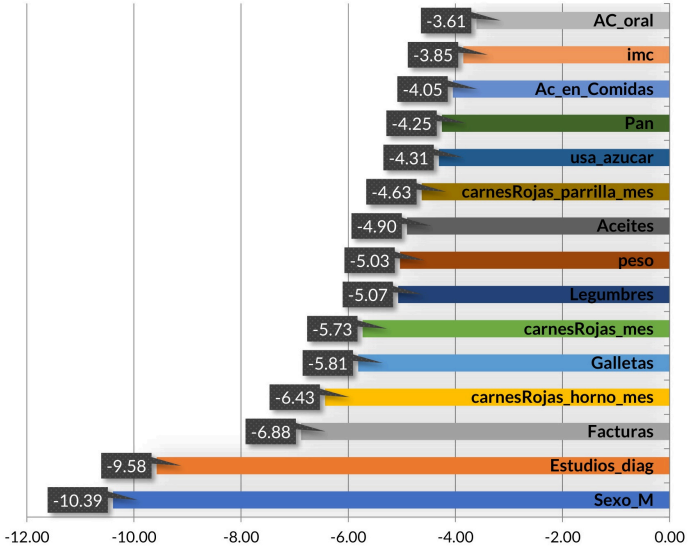
Figura 27. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de mama



La correlación negativa entre las variables con el cáncer de mama muestra que las de mayor importancia son: *sexo masculino*, *estudios diagnósticos*, *estatura*, *uso de anticonceptivos orales*, *consumo de aceites*, *azúcar*, *uso de aceites en comidas*, *consumo de carnes rojas al horno*, *facturas* y *peso*.

Entre las variables finales del *ranking* están: *realización de ejercicio físico al mes*, *no endulzar las infusiones*, *consumo de carnes blancas al horno por mes* y *consumo de pan*.

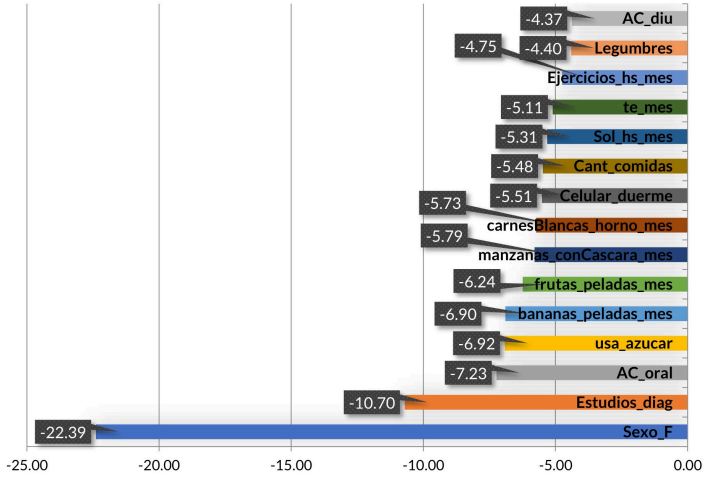
Figura 28. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de cuello de útero



La correlación negativa entre las variables con el cáncer de cuello de útero muestra que las de mayor importancia son: *sexo masculino*, *estudios diagnósticos*, *peso*, *consumo de facturas*, *legumbres*, *carnes rojas al horno al mes* y *peso*.

Entre las variables finales del *ranking* están: *consumo de aceites*, *carnes rojas a la parrilla por mes*, *uso de azúcar*, *pan*, *aceites en las comidas*, *uso de anticonceptivo orales* e *IMC*.

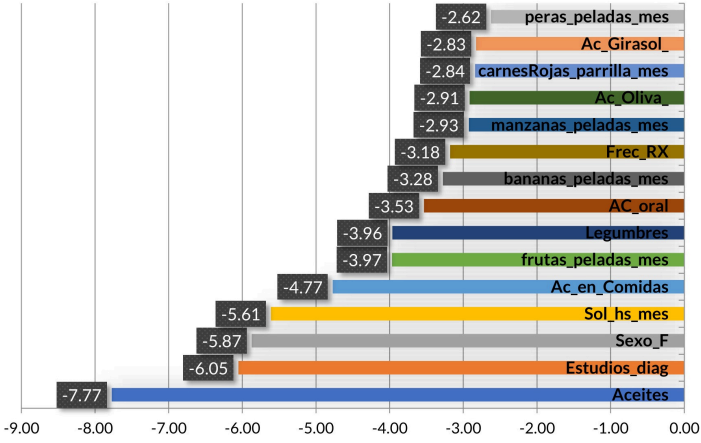
Figura 29. Gráfico de distribución de frecuencia según la correlación negativa del cáncer de próstata



La correlación negativa entre las variables con el cáncer de próstata muestra que las de mayor importancia son: *estudios diagnósticos, consumo de azúcar, frutas, bananas peladas y manzanas con cáscara al mes, carnes blancas al horno por mes, cantidad de comidas, consumo de té al mes, dormir con el celular* y *horas de exposición al sol por mes.*

Entre las variables finales del *ranking* están: *ejercicio físico en horas al mes, consumo de legumbres* y *uso de DIU.*

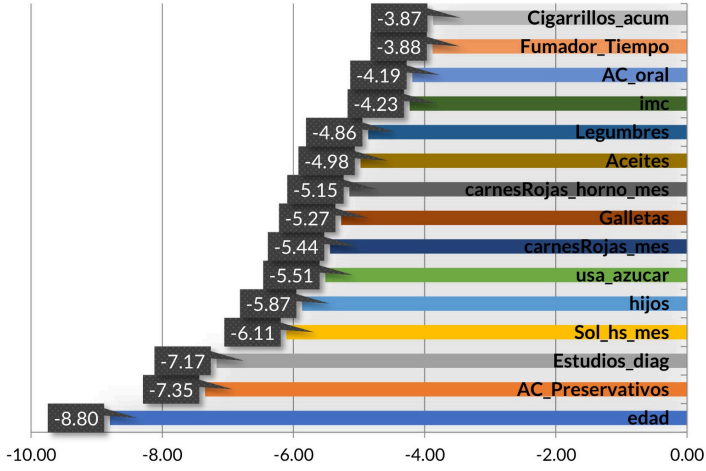
Figura 30. Gráfico de distribución de frecuencia según la correlación negativa del cáncer linfoma



La correlación negativa entre las variables con el linfoma muestra que las de mayor importancia son: *consumo de aceite y en las comidas, estudios diagnósticos, sexo femenino, horas de exposición al sol por mes, frutas y bananas peladas al mes, legumbres, uso de anticonceptivos orales* y *frecuencia de radiografías.*

Entre las variables finales del *ranking* están: *consumo de manzanas y peras peladas al mes, aceite de oliva y de girasol, consumo de carnes rojas a la parrilla al mes.*

Figura 31. Gráfico de distribución de frecuencia según la correlación negativa de leucemia



La correlación negativa entre las variables con la leucemia muestra que las de mayor importancia son: *edad*, *uso de preservativos*, *estudios diagnósticos*, *exposición al sol en horas al mes*, *número de hijos*, *consumo de azúcar*, *galletas* y *carnes rojas al horno por mes.*

Entre las variables finales del *ranking* están: *consumo de aceites y legumbres*, *IMC*, *uso de anticonceptivos orales*, *tiempo de fumador* y *cantidad acumulada de cigarrillos.*

4.2. Algoritmo de agrupamiento (clúster)

Al conjunto de datos se lo agrupa según criterios de distancia (acorde con el centroide).

- ***K-means***

Se divide el conjunto de datos en *k* grupos. Cada grupo se asocia a un centroide y los datos se corresponden con el centroide más cercano. Cada centroide representa el promedio de los elementos de cada uno de los *k* grupos (de ahí el nombre de la variable "*k-means*").

Este algoritmo no hace ningún tipo de análisis respecto de los casos de diagnóstico positivo o negativo de cáncer, sino que, simplemente, agrupa a los pacientes en *k* grupos según diferentes pares de variables elegidas. El valor agregado que le da interés a este trabajo está en intentar superponer los diagnósticos positivos con alguno de los *k* grupos, en caso de ser posible.

Para que el resultado tenga mayor interés y valor semántico, los pacientes se dividen en k=4 grupos. En general, cada grupo corresponderá con valores bajos o altos en una y otra variable, respectivamente.

Posteriormente, se busca que, para cada tipo de cáncer, todos los casos correspondientes se agrupen en una única categoría. De este modo, los resultados serán del tipo: "Todos los casos diagnosticados como positivos para el cáncer 'x' presentan un valor bajo en el atributo 'y' y alto en el atributo 'z'".

En el caso de variables dicotómicas, en lugar de emplear los términos "alto" o "bajo", resulta más adecuado utilizar "sí" o "no", según corresponda.

En algunos casos, se visualiza que alguna variable presenta valor nulo, sin ser dicotómica y, por ello, se coloca "no" en lugar de "bajo", ya que ofrece un resultado más enriquecedor.

Para la selección de variables en el armado de clústeres, se prueban todas las combinaciones de pares de variables que fueron seleccionadas según los criterios

expuestos en el capítulo anterior. El resultado se presenta siguiendo este mismo criterio: primero, se mostrarán los resultados de *clustering* con las variables seleccionadas por entropía, luego aquellos basados en la correlación positiva y, finalmente, los de correlación negativa. En cada caso, se presentarán combinaciones destacables para cada tipo de cáncer que se analiza, siempre que alguna combinación cumpla los requisitos previamente establecidos.

- ***K-means* con variables seleccionadas con entropía**

A continuación, se presenta un listado por tipo de cáncer, indicando si existen combinaciones de variables seleccionadas por entropía de Shannon, donde todos los casos diagnosticados como positivos están totalmente incluidos dentro de un mismo tipo de cáncer.

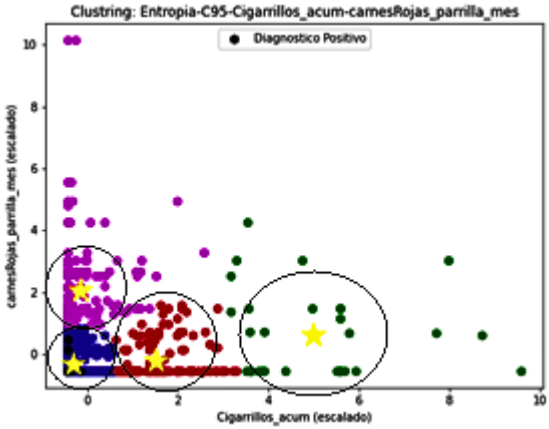
La separación en cuatro grupos se realizó con base en los siguientes criterios:

a. Proximidad entre los datos (criterios de cercanía).
b. Valores altos o bajos en dos variables, lo que da como resultado las cuatro combinaciones o grupos posibles, representados en la gráfica con los colores verde, rojo, azul y fucsia. Estos colores son asignaciones jerárquicas y ejemplifican los valores agrupados por el algoritmo, representan una escala nominal.

Esta clasificación permite una mayor especificidad en el análisis. No se optó por dividir en dos o tres grupos porque la información obtenida era trivial; por ejemplo, en el caso del cáncer de próstata, todos los datos recaerían en el sexo masculino.

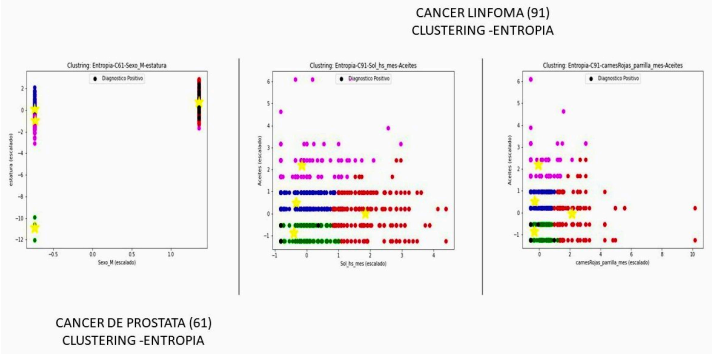
En las gráficas, los puntos negros representan los casos de este tipo de cáncer incluidos en la muestra, mientras que las estrellas señalan el centroide de cada grupo (centro de masa), es decir, la media de los datos (promedio).

Figura 32. ***Clustering*** **de diversas variables**



En las gráficas correspondientes al cáncer de próstata y linfoma (entropía), se observa la agrupación de datos basada en medidas de distancia, específicamente la distancia euclidiana o norma euclidiana (la distancia entre dos puntos de una gráfica). Además, los datos están escalados, lo que significa que han sido normalizados para garantizar que la distancia euclidiana sea coherente.

Figura 33. ***Clustering*** **para estudiar cáncer de próstata y linfoma**



La dicotomía "poco" o "mucho" para una variable indica que, del 100 % de los casos de todos los tipos de cáncer, se puede determinar a qué grupo pertenece cada uno. Esto significa que, para un tipo específico de cáncer, todos los pacientes comparten valores categorizados como "poco" o "mucho" respecto a una variable determinada. Sin embargo, esta clasificación no implica causalidad; simplemente es una característica descripta observada en los datos.

Para algunos cánceres, esta herramienta se encontró con ciertas variables que no pudo encasillar, por lo que simplemente las eliminó. Por ejemplo, en el C-34 (cáncer de pulmón) desestimó la variable "cigarrillos acumulados", ya que esta variable estaba tanto en el grupo de los fumadores activos como en el de fumadores pasivos.

Otra observación a partir de las tablas con los resultados (ver tablas por diagnóstico en Anexo III) muestra que todos los casos positivos de cáncer tienen siempre "poco" de las dos variables combinadas por *clustering*. Lo mencionado anteriormente se aplica tanto a la entropía como a la correlación positiva y negativa.

La separación en cuatro grupos aleatorios planteados por la herramienta *k-means* acorde con los criterios de cercanía muestra que la dispersión de los cánceres siempre da "pocos" entre las variables seleccionadas y que esto se atribuye al número de casos analizados. Como se pensó que este resultado del primer análisis podía ser un sesgo, se eliminaron las restricciones al algoritmo (sin límite en la toma de datos). Sin embargo, no se registraron cambios significativos.

Se puede concluir que este tipo de algoritmo no aporta información relevante, por lo que no se aplicó en el segundo análisis.

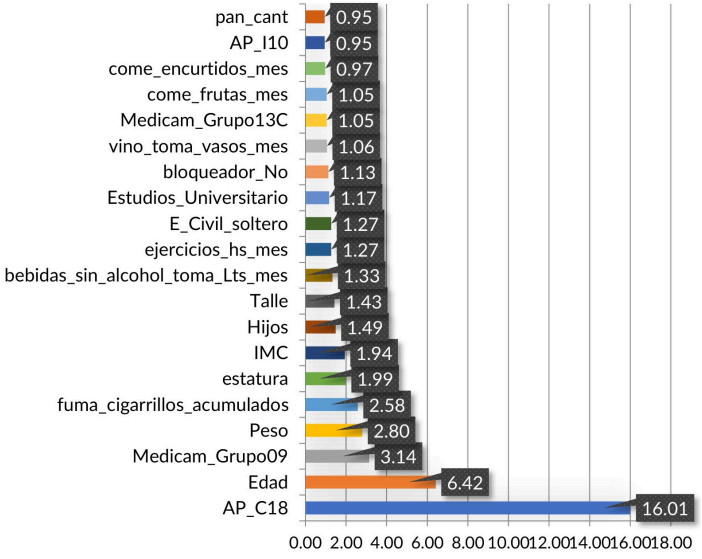
Segundo análisis

4.3. Reducción de dimensionalidad

4.3.1. Entropía de Shannon

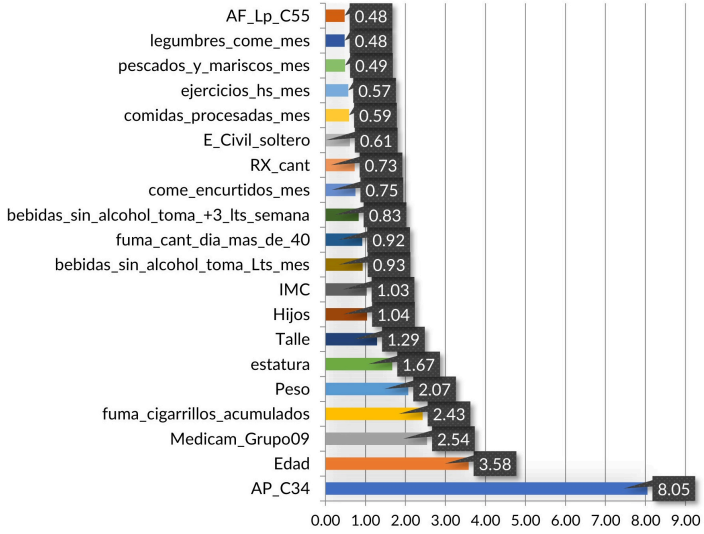
Se presenta un *ranking* con las veinte variables que más contribuyen en términos de información a las variables objetivo asociadas con los diferentes diagnósticos de cáncer.

Figura 34. Gráfico de distribución de frecuencia según la entropía del cáncer de colon



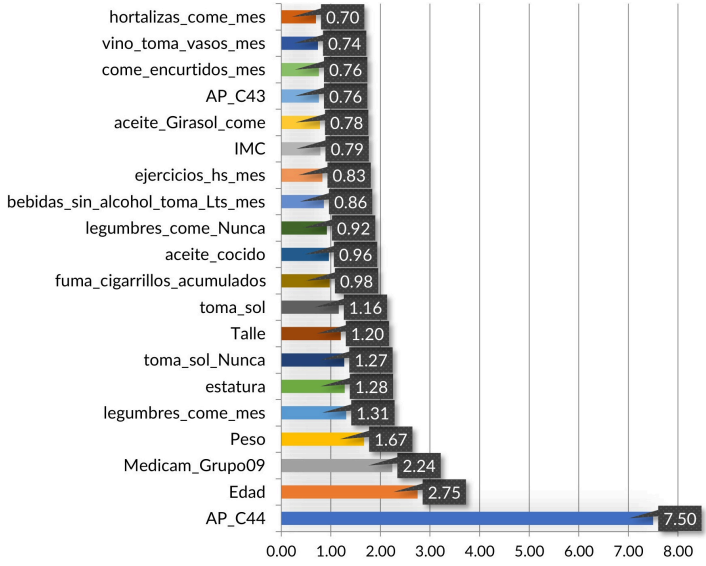
Para el cáncer de colon las variables con mayor relación en entropía son: *consumo acumulado de cigarrillos*, *consumo de bebidas sin alcohol en litros por mes*, seguido por *cantidad de ejercicios en horas por mes.* Asimismo, se observa *no uso de bloqueadores solares, consumo de vasos al mes de vino.* Por último, se observa que las últimas cinco variables de menor influencia para este cáncer son: *uso de medicamentos hipertensivos, consumo al mes de frutas, consumo de encurtidos, cantidad de pan* y *antecedente personal de sufrir hipertensión arterial.*

Figura 35. Gráfico de distribución de frecuencia según la entropía del cáncer de pulmón



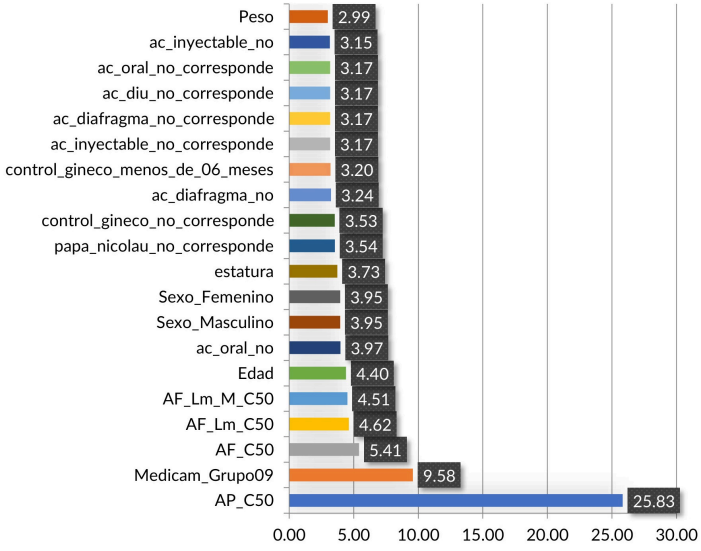
En el cáncer de pulmón las variables que más se relacionan con la entropía son: *consumo acumulado de cigarrillos, consumo de bebidas sin alcohol en litros al mes, consumo de cigarrillos (más de 40 al día).* Entre las variables menos influyentes se encuentran: *cantidad de radiografías, consumo de bebidas sin alcohol (más de 3 litros por semana), consumo de encurtidos al mes, consumo de comidas procesadas al mes, consumo de pescados, mariscos y legumbres al mes, ejercicio físico al mes.* Finalmente aparece *antecedente familiar, en la línea paterna, del cáncer de útero*.

Figura 36. Gráfico de distribución de frecuencia según la entropía del cáncer de piel



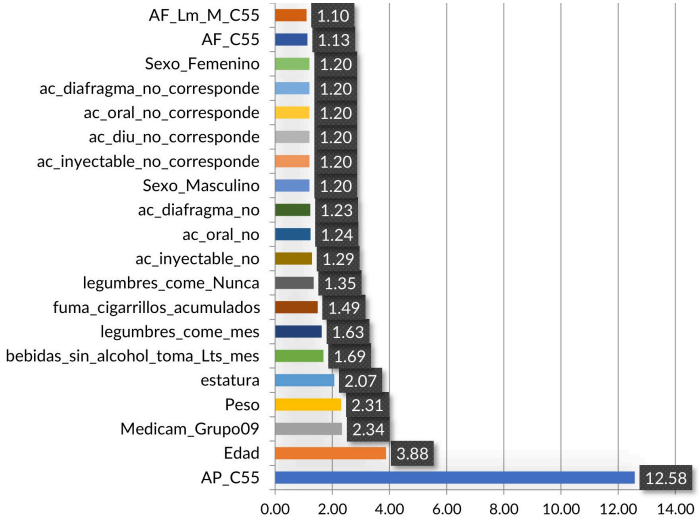
En el análisis para el cáncer de piel, las variables con mayor correlación son: *nunca toma sol* y *quienes sí toman sol* seguidas por *consumo acumulado de cigarrillos.* Entre las variables de hábitos alimentarios están: *consumo de legumbres al mes, consumo de bebidas sin alcohol en litros al mes, consumo de aceite de girasol, encurtidos, hortalizas al mes, consumo de vino en cantidad de vasos por mes.* Además, se destacan las variables: *ejercicio físico en horas por mes* y *antecedente personal de melanoma.*

Figura 37. Gráfico de distribución de frecuencia según la entropía del cáncer de mama



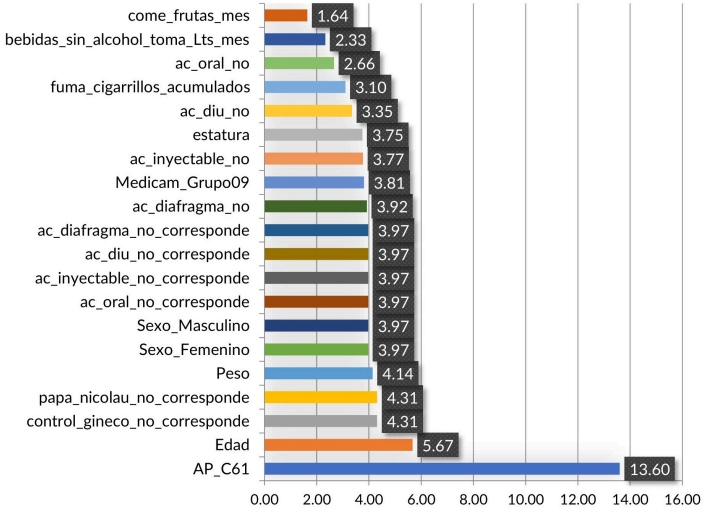
El cáncer de mama se correlaciona con mayor intensidad con las siguientes variables: *antecedentes familiares, por la línea materna, en mujeres.* Luego siguen en orden de importancia*: no usa anticonceptivo oral*, con igual peso en el sexo femenino y el masculino, y *control ginecológico cada 6 meses.*

Figura 38. Gráfico de distribución de frecuencia según la entropía del cáncer de cuello de útero



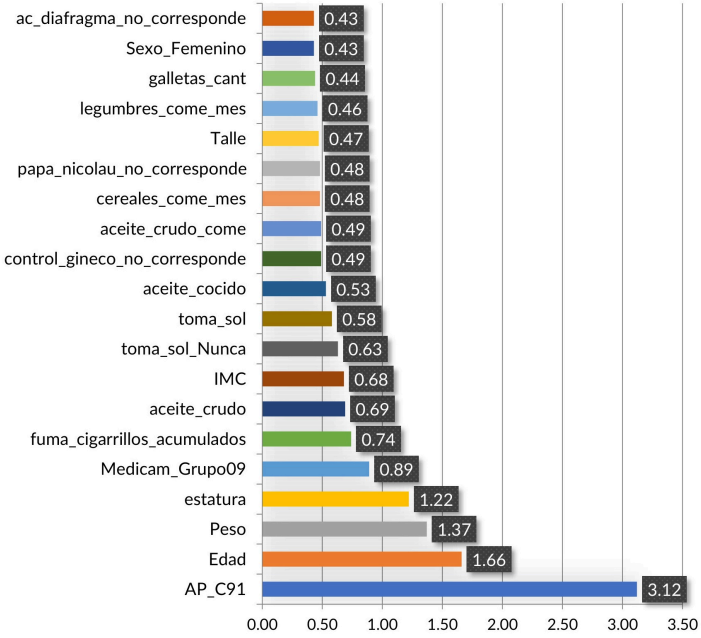
El cáncer de cuello de útero se correlaciona con mayor intensidad con las variables siguientes: *consumo de bebidas sin alcohol en litros por mes, legumbres al mes y que nunca consume, consumo acumulado de cigarrillos.* Luego siguen en orden de relevancia las variables*: no usa anticonceptivo oral y diafragma.* Al final de este *ranking* aparecen: *antecedentes familiares, línea materna, mujer con este cáncer.*

Figura 39. Gráfico de distribución de frecuencia según la entropía del cáncer de próstata



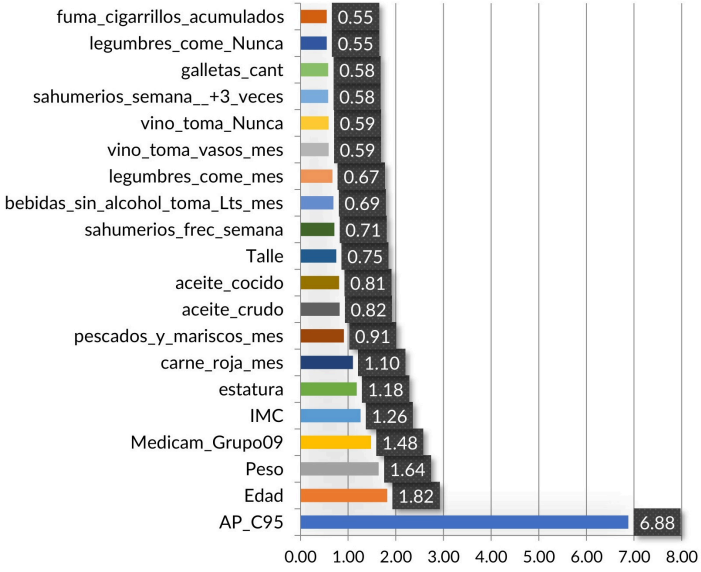
Para el cáncer de próstata, se desestiman las variables que se consideran tautológicas, sin embargo, las últimas variables del *ranking* muestran asociación con este cáncer. Estas son: *consumo acumulado de cigarrillos, consumo de bebidas sin alcohol en litros al mes* y *consumo de frutas al mes*.

Figura 40. Gráfico de distribución de frecuencia según la entropía del diagnóstico de linfoma



Para el linfoma, las variables con más correlación son: *consumo acumulado de cigarrillos* y *de aceite crudo.* Seguido por *exposición al sol, consumo de aceite cocido, aceite crudo en las comidas, cereales y legumbres que consume por mes, cantidad de galletas, sexo femenino.* De igual manera están las variables que no se corresponden, como: *control ginecológico, Papanicolaou, anticonceptivo diafragma.*

Figura 41. Gráfico de distribución de frecuencia según la entropía del diagnóstico de leucemia

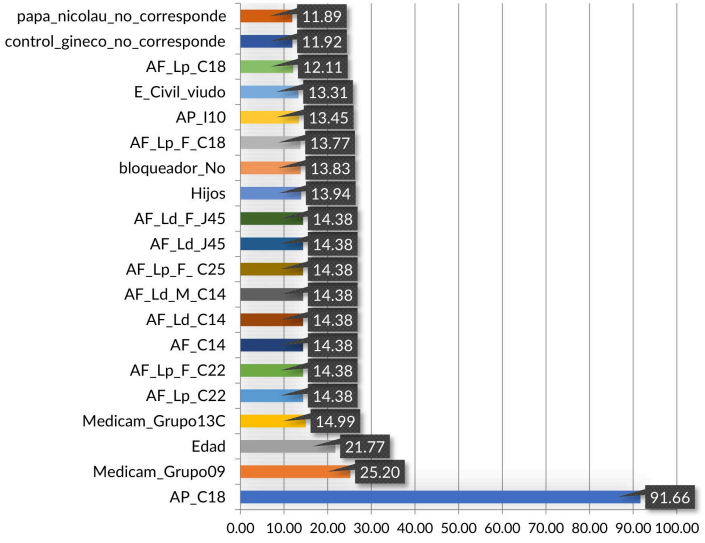


La leucemia es un diagnóstico que se relaciona con mayor frecuencia con las siguientes variables: *consumo de carnes rojas, pescados y mariscos al mes, consumo de aceite crudo y cocido*. Le siguen en orden de relevancia: *uso frecuente de sahumerios en la semana, consumo de bebidas sin alcohol en litros al mes, consumo de legumbres al mes, consumo de vino en vasos o no al mes, cantidad de galletas* y, por último, *consumo acumulado de cigarrillos.*

- **Análisis por correlación positiva**

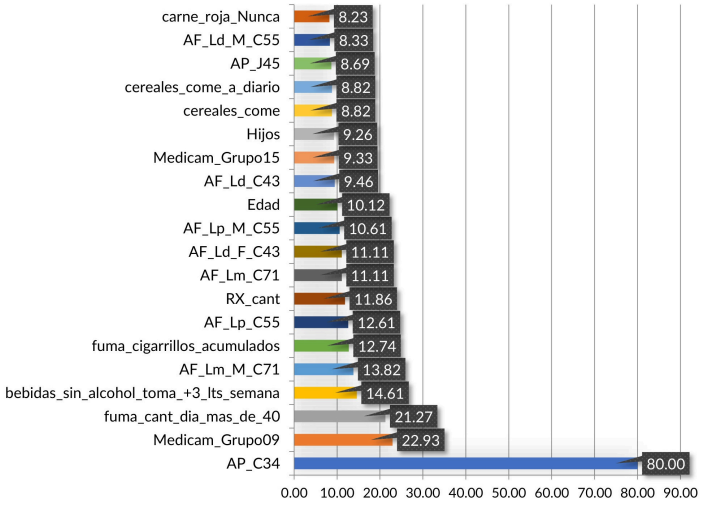
Se presentan las veinte variables con la mejor correlación positiva con el diagnóstico positivo de cada tipo de cáncer.

Figura 42. Gráfico de distribución de frecuencia según correlación positiva del cáncer de colon



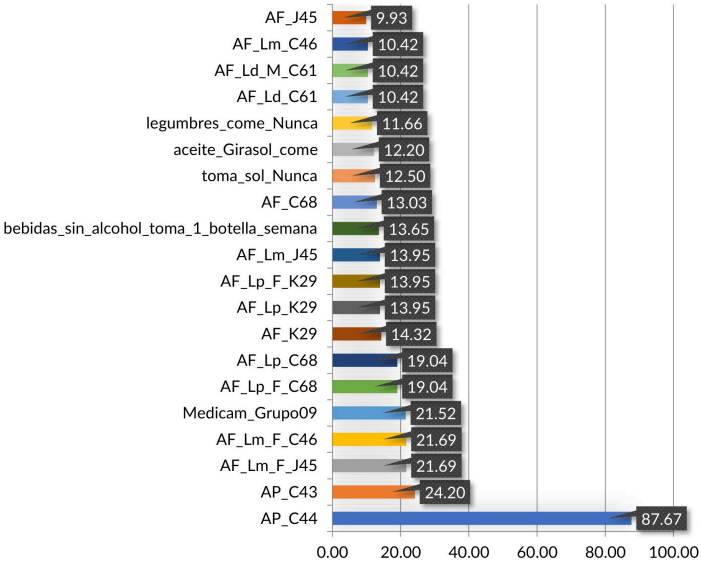
La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *edad, uso de medicamentos antihipertensivos arteriales, antecedente familiar, línea paterna, de cáncer de hígado, páncreas, colon, en la línea directa, cáncer de cabeza y cuello, asma, antecedente personal de hipertensión arterial* y *no uso de bloqueadores solares.*

Figura 43. Gráfico de distribución de frecuencia según correlación positiva del cáncer de pulmón



La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *edad, fuma más de 40 cigarrillos al día, cantidad de Rx, antecedente familiar, línea materna, de cáncer de encéfalo, cuello de útero, melanoma, medicamentos antihiperglucemiantes, consumo de cereales, consumo de bebidas sin alcohol (más de 3l a la semana).*

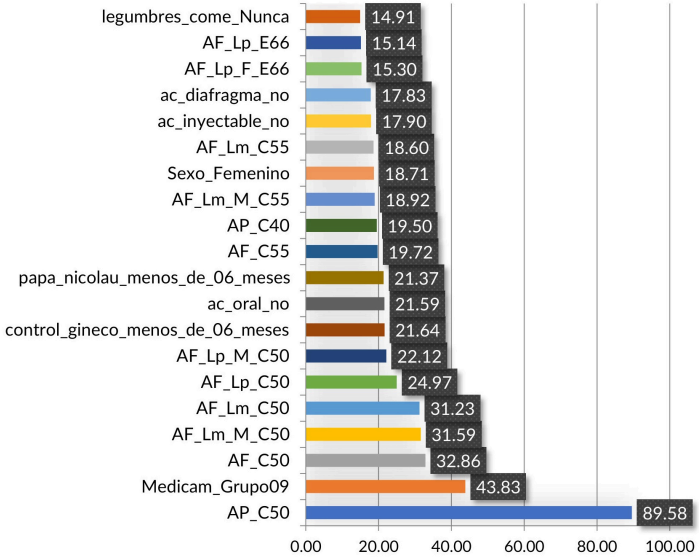
Figura 44. Gráfico de distribución de frecuencia según correlación positiva del cáncer de piel



La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *antecedentes familiares en la línea materna y paterna para el melanoma, asma, sarcoma de Kaposi, cáncer de riñón, gastritis crónica, consumo de bebidas sin alcohol (una botella a la semana), no toma sol, no consume aceite de girasol y legumbres.*

Entre las variables finales del *ranking* están: *antecedentes familiares en la línea directa y materna para cáncer de próstata, sarcoma de Kaposi y asma.*

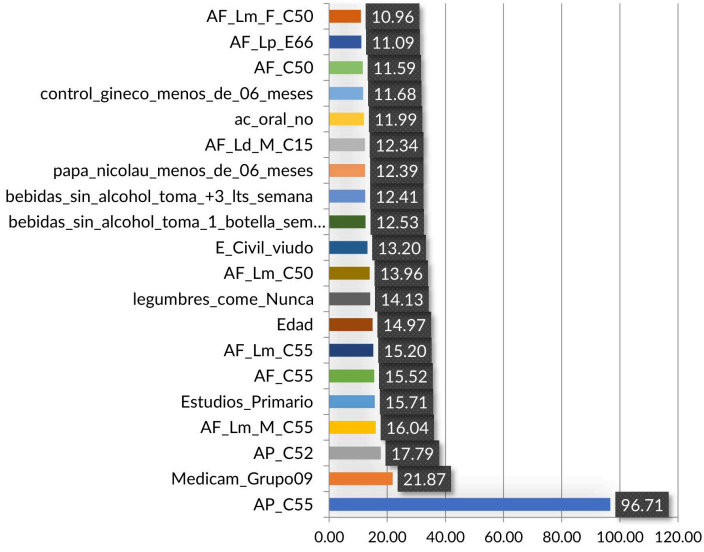
Figura 45. Gráfico de distribución de frecuencia según correlación positiva del cáncer de mama



La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *antecedente familiar de cáncer de mama por línea materna y paterna, no realizarse controles ginecológicos y Papanicolaou cada 6 meses, no uso de anticonceptivos orales, inyectables y diafragma* y *antecedentes familiares de cáncer de cuello de útero.*

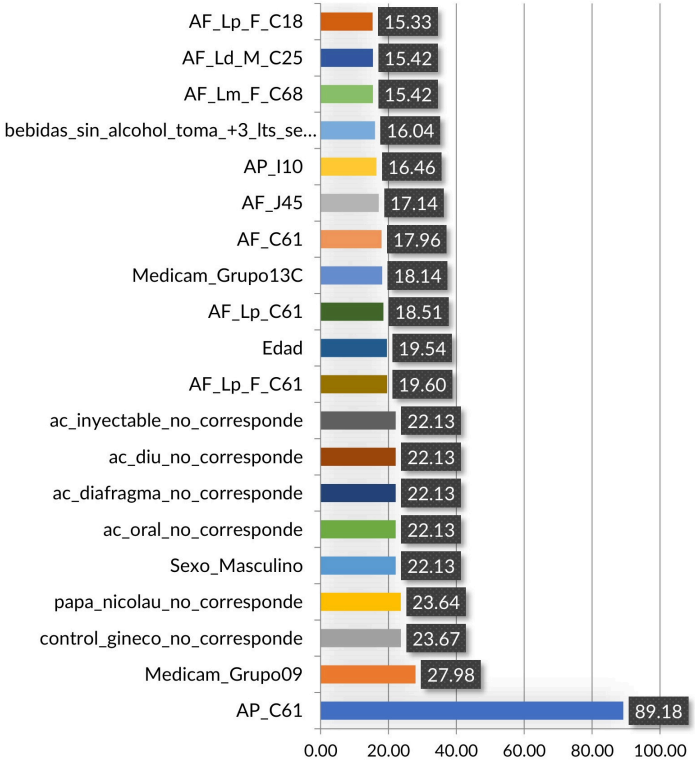
Entre las variables finales del *ranking* están: *antecedentes familiares de obesidad en la línea paterna* y *no consumo de legumbres.*

Figura 46. Gráfico de distribución de frecuencia según correlación positiva del cáncer de cuello de útero



La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *antecedentes personales de tumor maligno en vagina, antecedente familiar, línea materna, de cáncer de cuello de útero y mama, nivel de estudios completo primaria, edad, no consumo de legumbres, estado civil viudo, consumo de bebidas sin alcohol (más de 3 litros a la semana y en botella), no realizarse controles de ginecología y Papanicolaou cada 6 meses.* Entre las variables finales del *ranking* están: *antecedente familiar, línea directa, de cáncer de esófago, no uso de anticonceptivos orales* y *antecedente familiar, línea paterna, de obesidad.*

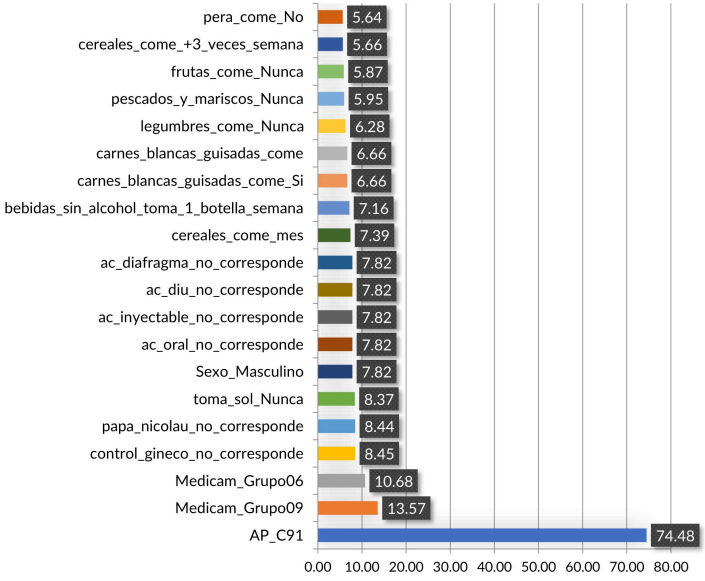
Figura 47. Gráfico de distribución de frecuencia según correlación positiva del cáncer de próstata



La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *antecedente familiar, línea paterna, de cáncer de próstata, edad, uso de medicamentos hipertensivos arteriales, antecedentes familiares de hipertensión arterial y asma, consumo de más de 3 litros de bebidas sin alcohol a la semana.*

Entre las variables finales del *ranking* están: *antecedentes familiares, línea materna, cáncer de riñón, línea directa, cáncer de páncreas* y *línea paterna, cáncer de colon.*

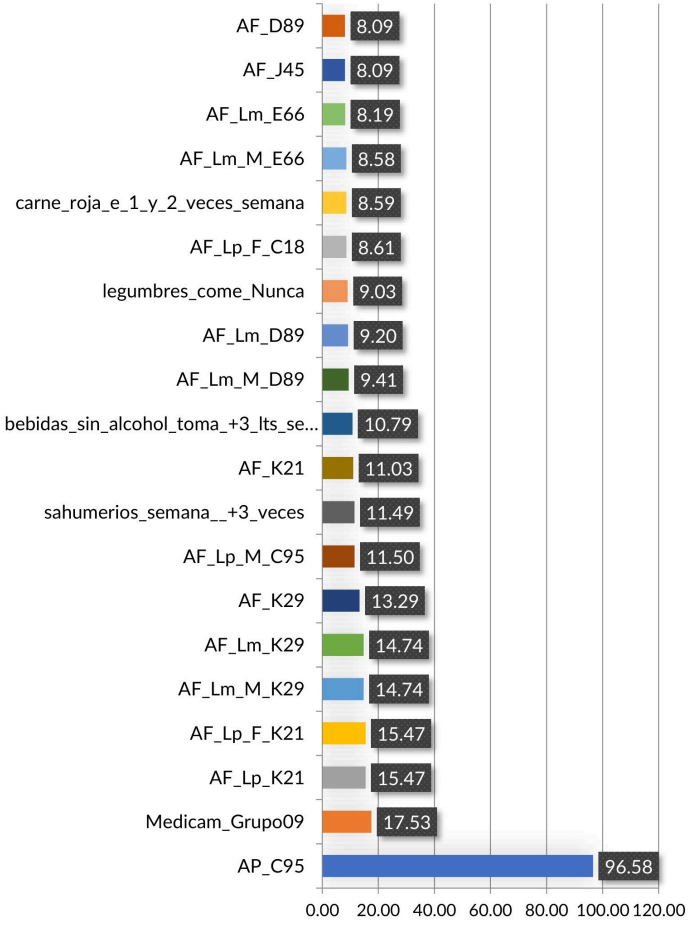
Figura 48. Gráfico de distribución de frecuencia según correlación positiva del diagnóstico de linfoma



La correlación positiva entre las variables con la variable objetivo nos permite ver que las de mayor importancia son: *no exposición a la luz solar, consumir cereales al mes, consumo semanal de bebidas sin alcohol, consumo de carnes blancas en guiso, no consumo de legumbres, pescados y mariscos.*

Entre las variables finales del *ranking* están: *no consumir frutas* y *comer cereales más de 3 veces a la semana.*

Figura 49. Gráfico de distribución de frecuencia según correlación positiva del diagnóstico de leucemia



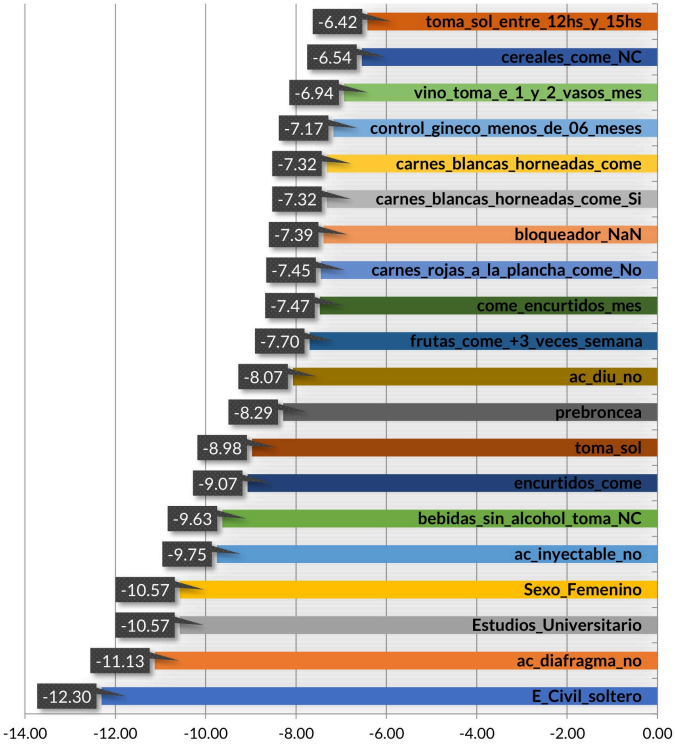
La correlación positiva entre las variables con la variable objetivo muestra que las de mayor importancia son: *antecedente familiar de reflujo gastroesofágico, cáncer de colon y leucemia por línea paterna, antecedentes de gastritis crónica y celiaquía en la línea materna, uso de sahumerios más de 3 veces a la semana, consumo de bebidas sin alcohol (más de 3 litros a la semana)* y *no consumir legumbres.*

Entre las variables finales del *ranking* están: *consumo de carnes rojas entre 1 y 2 veces por semana, antecedentes familiares de celiaquía, asma y de obesidad por línea materna.*

- **Análisis por correlación negativa**

Se presentan las veinte variables que mejor se correlacionan negativamente con el diagnóstico positivo de cada tipo de cáncer.

Figura 50. Gráfico de distribución de frecuencia según correlación negativa del cáncer de colon

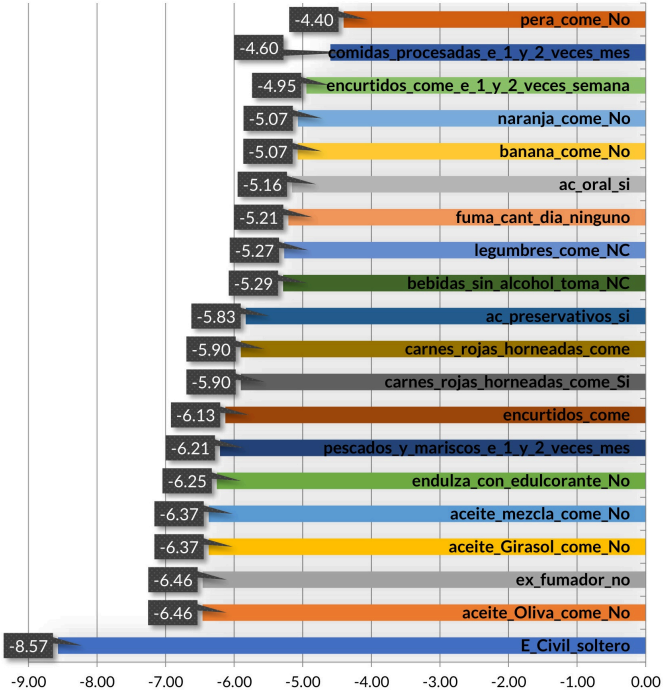


Nota. El dato demográfico "estado civil soltero" muestra que hay preponderancia de sujetos solteros.

La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *sexo femenino, no uso de métodos anticonceptivos (diafragma, inyectables, DIU), se realiza controles ginecológicos en intervalos menores a 6 meses, no contesta sobre el consumo de bebidas sin alcohol, exposición a la luz solar entre las 12 h y las 15 h y prebronceado, consume encurtidos, consume*

frutas más de 3 veces a la semana, consume carnes blancas horneadas, consume cereales, no consume carnes rojas a la plancha.

Figura 51. Gráfico de distribución de frecuencia según correlación negativa del cáncer de pulmón

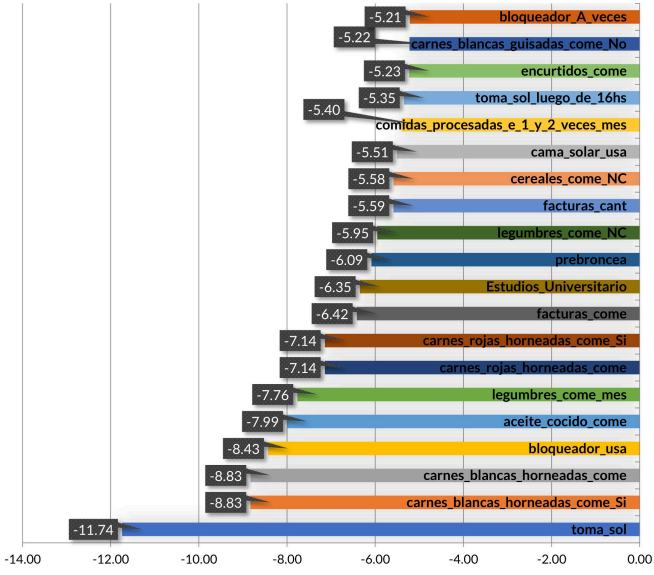


Nota. El dato demográfico "estado civil soltero" solo muestra que quienes respondieron eran solteros en su mayoría.

La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *no haber sido fumador, uso de preservativos y anticonceptivos*

orales, no consumir aceite de oliva, girasol y mezclas, no usar edulcorante, no comer bananas, naranjas y peras, no responde sobre consumo de legumbres. Adquieren mayor relevancia para este análisis las variables: *consumir comidas procesadas, pescados y mariscos entre 1 y 2 veces al mes, encurtidos de 1 a 2 veces a la semana* y *consumir carnes rojas al horno.*

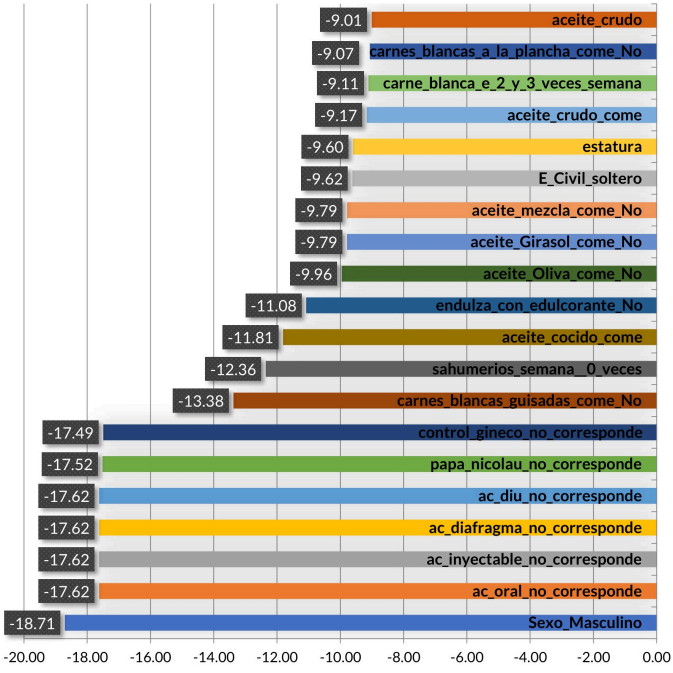
Figura 52. Gráfico de distribución de frecuencia según correlación negativa del cáncer de piel



La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *estudios universitarios, exposición a la luz solar (después de las 16 h), prebronceado, uso de cama solar, uso de bloqueador a veces, consumo de carnes blancas y rojas al horno,*

consumo de comidas procesadas, encurtidos y facturas entre 1 y 2 veces al mes, no contesta sobre consumo de cereales y legumbres, no consume carnes blancas guisadas.

Figura 53. Gráfico de distribución de frecuencia según correlación negativa del cáncer de mama



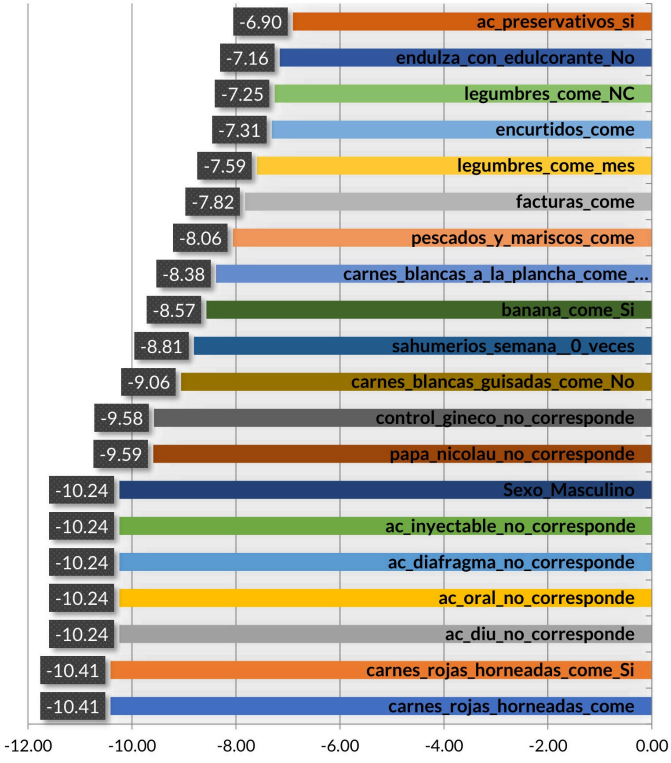
Nota. El dato demográfico "estado civil soltero" solo muestra que quienes respondieron eran solteros en su mayoría.

La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *sexo masculino, estatura, no usa sahumerios.* No

corresponde: *uso de anticonceptivos orales (inyectables, diafragma, DIU), control ginecológico cada 6 meses* y *estudio de Papanicolaou.*

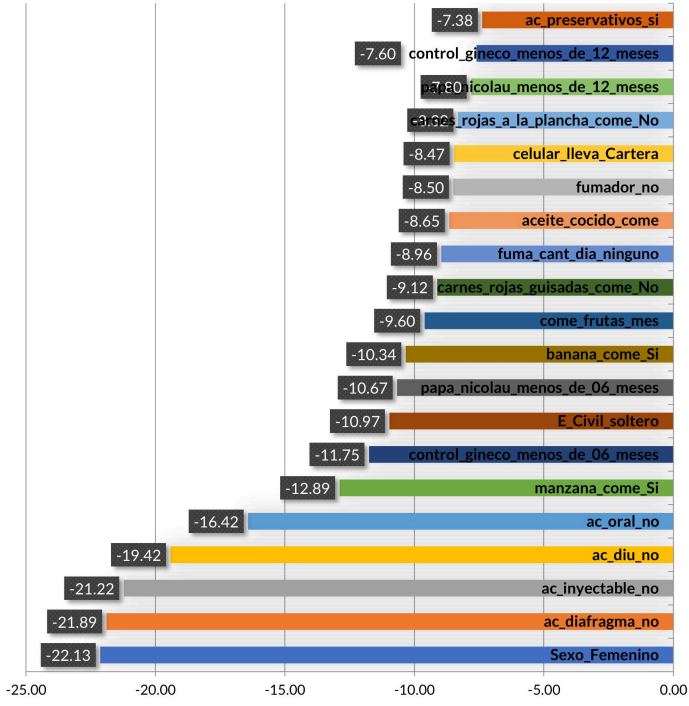
Las variables que podrían ser mayores predisponentes al cáncer son: *no consumo de carnes blancas guisadas y a la plancha, ni aceite de girasol, oliva y mezcla, consumo de carnes blancas entre 2 y 3 veces a la semana.*

Figura 54. Gráfico de distribución de frecuencia según correlación negativa del cáncer de cuello de útero



La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *sexo masculino, uso de preservativos, consumo de carnes rojas al horno, pescados y mariscos, consumo de facturas, encurtidos y bananas, no consumo de carnes blancas a la plancha, legumbres* y *no uso de edulcorante*.

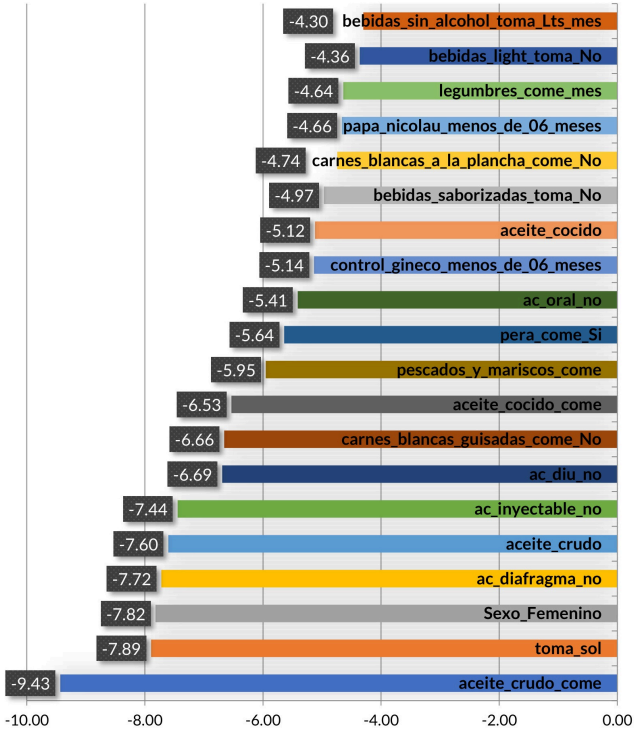
Figura 55. Gráfico de distribución de frecuencia según correlación negativa del cáncer de próstata



Nota. El dato demográfico "estado civil soltero" solo muestra que quienes respondieron mayormente eran solteros.

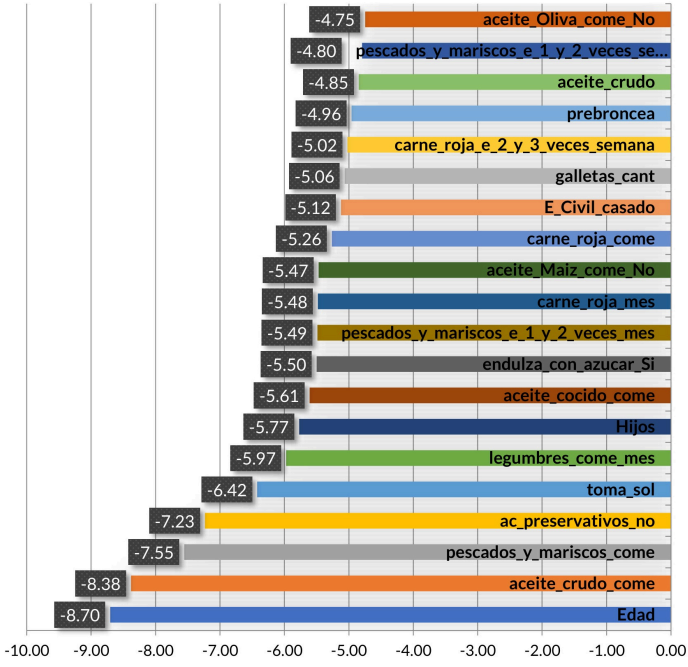
La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *sexo femenino, no uso de anticonceptivos diafragma, DIU, inyectable, oral, pero sí uso de preservativo, se realiza control ginecológico cada 6 y 12 meses o menos* y *se realiza el estudio Papanicolaou, no haber sido fumador, uso del celular en la cartera, consumo de manzanas, bananas, frutas, consumo de aceite cocido, no consumo de carnes rojas guisadas.*

Figura 56. Gráfico de distribución de frecuencia según correlación negativa del diagnóstico de linfoma



La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *sexo femenino, consumo de aceite crudo y cocido, consumo de legumbres, peras, pescados y mariscos, bebidas sin alcohol en litros al mes, no consumo de bebidas light, carnes blancas guisadas y a la plancha, no uso de anticonceptivo DIU, inyectable, diafragma y oral.*

Figura 57. Gráfico de distribución de frecuencia según correlación negativa del diagnóstico de leucemia



Nota. El dato demográfico "estado civil casado" solo muestra que quienes respondieron mayormente eran casados.

La correlación negativa entre las variables con la variable objetivo muestra que las de mayor importancia son: *edad, hijos, exposición a la luz solar, prebronceado, consumo de aceite crudo y cocido, aceite de maíz, consumo de galletas, pescados y mariscos 1 a 2 veces a la semana, legumbres, carnes rojas 2 a 3 veces a la semana, endulza con azúcar, no consumo de aceite de oliva.*

5

A modo de cierre

Este trabajo conjuga la integración de datos multiómicos, es decir, aquellos que se centran en el estudio de las interacciones de los distintos factores que definen a un sujeto, tales como su genética y su epigenética. En la elaboración de los resultados se recurrió al uso y desarrollo de la IA, considerando estos dos parámetros: a) el enfoque de la medicina frente a grandes y complejos conjuntos de datos junto con las herramientas informáticas necesarias para procesarlos en el menor tiempo posible; b) la identificación de patrones que permitiesen obtener clasificaciones o predicciones con mayor precisión. Al momento de realizar ensayos clínicos, el tiempo que se pierde en la codificación de los datos médicos es muy grande. Las herramientas de la IA, por su parte, permiten acelerar y obtener una pesquisa más rápida e inteligente de aquellos datos nuevos o de mayor importancia.

Para el uso de la IA se deben respetar los siete principios éticos[12] que se mencionan a continuación: 1) intervención y supervisión humana en su aplicación y uso, 2) garantía de solidez y seguridad técnica, 3) protección en la privacidad y gestión de los datos, 4) transparencia, 5) diversidad y no discriminación, 6) bienestar social y medioambiental, y 7) rendición de cuentas.

---

12 Comisión Europea (2018), https://tinyurl.com/2t3aac4m.

Se partió de un universo con 400 casos y 2200 controles. Se utilizó un cuestionario con 849 variables y se evaluaron los resultados mediante *machine learning* y algoritmos de aprendizaje supervisado y no supervisado. El objetivo fue encontrar las variables con mayor asociación al riesgo de desarrollar enfermedad neoplásica. Posteriormente, la herramienta Azure de Microsoft identificó ocho tipos de cáncer con mayor influencia en función de la cantidad de casos en la muestra, los cuales se denominaron variables *target* u objetivo. Esta selección no se pudo hacer directamente con las encuestas, debido al gran número de encuestadores presenciales que participaron en la recolección de los datos.

Para ello, en una primera instancia se utilizaron ocho variables objetivo versus 88 variables independientes, se omitieron los antecedentes familiares y la medicación (véase listado 1 en Anexo II). Por otro lado, en un segundo momento, se incluyeron las variables relacionadas con estos ítems. El análisis se hizo entre ocho variables objetivo versus 849 variables independientes (véase listado 2 en Anexo II).

Para lograr un modelo adecuado que permita obtener patrones o clasificaciones verdaderas, se deben etiquetar correctamente los datos del conjunto. Al no haber un trabajo previo similar en cuanto al conjunto de datos, estos se tuvieron que "normalizar o codificar". De igual manera se tuvo que proceder con los valores para las variables dependientes y armar los "dataset" necesarios para que los algoritmos de la IA pudiesen procesarlos. Se utilizaron protocolos ya estandarizados por la Organización Mundial de la Salud para codificar los cánceres, enfermedades (CIE-10) y los medicamentos.

En agosto de 2022, se obtuvieron los resultados del primer análisis. El estudio de los resultados llevó a ampliar el conjunto de datos, codificando los apartados antecedentes personales y medicación, para evaluar si los nuevos cruces de variables arrojarían nuevos datos con mayor correlación entre las variables objetivo y las independientes.

Cuando se diseñó este trabajo poco se conocía acerca de los análisis por aprendizaje automático en medicina; sin embargo, en un tiempo ulterior esa información fue evolucionando y, actualmente, las publicaciones sobre este tipo de procesamiento son variadas. A pesar de ello, todas estas están fundadas en bases de datos públicas. Algunas de ellas, disponibles para estudios de cáncer, son: *The Cancer Genome Atlas* (TCGA), el banco de tumores de Rotterdam, Taxonomía Molecular del Consorcio Internacional del Cáncer de Mama (METABRIC, por su sigla en inglés), el conjunto de datos MITOS-ATYPIA-14, *Tumor Proliferation Assessment Challenge* (TUPAC) 2016 dataset, INbreast database, *Lung Image Database Consortium and Image Database Resource Initiative* (LIDC-IDRI) datasets, *Lung Nodule Analysis* (LUNA16) dataset, *Breast Cancer Histopathological Image Classification* (BREASTHIs) dataset 2015, Bioimaging conjunto de datos de detección de metástasis de cáncer en ganglios linfáticos (CAMELYON, por su sigla en inglés), conjunto de datos PatchCamelyon, base de datos TCGA. Por otra parte, a nivel local no existen trabajos con este precedente (116, 119).

En el caso del aprendizaje supervisado, se puede entrenar al algoritmo a través de la carga de datos de los pacientes diagnosticados mediante un profesional de la salud y garantizar su validación. Esto permitirá que, ante nuevos datos, el modelo de aprendizaje automático los analice de manera independiente, lo que hace posible la comparación de resultados.

Gracias al desarrollo de este trabajo, el aprendizaje supervisado demostró la validez del entrenamiento, más allá de que los resultados no fueron del todo novedosos. En la información dada por el "Aprendizaje supervisado", pudo observarse que las resultantes del primer y segundo análisis fueron similares, por eso no se incluyeron en el presente trabajo. Este entrenamiento se utiliza para predicción y diagnóstico, y los mejores modelos son *super vector machine* y *random forest* (120, 121).

Los análisis de los datos de las encuestas mediante entropía y correlación positiva por aprendizaje no supervisado confirman los datos bibliográficos y, por lo tanto, validan el método de análisis utilizado. A pesar de lo anterior, se menciona que no han generado aportes de significancia al conocimiento sobre nuevos predictores de enfermedad neoplásica.

Estos resultados coinciden en que las variables *target* comparten las variables: *IMC*, *peso*, *estatura*, *edad*, *consumo de cigarrillos*, *consumo de carnes blancas y rojas* y los *antecedentes familiares*. Estos resultados coinciden con los obtenidos mediante aprendizaje supervisado.

Los dos análisis realizados mediante el modelo de correlación negativa por aprendizaje no supervisado aportaron datos relevantes, mostraron que aquellas acciones consideradas en un principio como preventivas del desarrollo de enfermedades neoplásicas en realidad también son factores predisponentes.

Es de resaltar que la correlación positiva da como resultados aquellas variables más afines a la variable objetivo, mientras que la correlación negativa muestra aquellas variables que se contraponen a las variables objetivo, es decir, las acciones que se realizan como contramedidas.

Se encontró que las siguientes variables se reiteran para cada variable objetivo: consumo de carnes rojas y blancas en cualquiera de su forma de cocción, y la cantidad que se consume de ellas; exposición solar; consumo de azúcar y edulcorantes, aceites, legumbres, encurtidos, frutas peladas; actividad física; y anticonceptivos orales. Estas variables aparecen en un doble rol, como protectores y como predisponentes para el desarrollo de enfermedad neoplásica. Se podrían explicar estos resultados desde la hormesis,[13] es decir, aquellos elementos que en bajas dosis tienen efectos protectores, mientras que en dosis altas producen efectos adversos.

Actualmente, el cáncer se considera una consecuencia de interacciones complejas entre el genoma y el ambiente. Las alteraciones genéticas y epigenéticas son dos mecanismos independientes que participan en su aparición y progresión. La desregulación epigenética de los procesos de transcripción y traducción de genes mutados, que están relacionados con el mecanismo del ciclo celular, da como resultado el crecimiento descontrolado de las células tumorales.

Los mecanismos epigenéticos regulan la expresión de los genes sin alterar las secuencias de bases del ADN. Esto establece un vínculo entre las influencias ambientales y genéticas que determinan el fenotipo. Además, las células desarrollan patrones epigenéticos especializados que son esenciales para su función y diferenciación.

---

13 Hormesis (del griego *hormaein* –estimular–, del prefijo *horm* deviene hormona) fue definida como la respuesta bifásica en que ciertos agentes químicos y físicos afectan a los seres vivos: dosis bajas provocan efectos "favorables", dosis altas provocan efectos "adversos". http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0025-76802005000100017.

Las histonas sufren numerosas modificaciones postradicionales, como metilación, desmetilación, acetilación, desacetilación, fosforilación, ubiquitinación.

El nivel de metilación, ya sea de tipo hipo o híper, está influenciado por factores internos y externos, tales como los hidrocarburos aromáticos policíclicos (HAP), un grupo amplio de sustancias químicas diferentes que se forman principalmente durante la combustión incompleta de materia orgánica. Algunos de estos tienen definidas sus propiedades carcinogénicas, mutagénicas y teratogénicas. Los HAP presentes en los alimentos representan un riesgo potencial para el consumo de alimentos contaminados. El benzo[a]pireno (BaP) es uno de los primeros carcinógenos descubiertos, y se lo encuentra por ejemplo en el humo del cigarrillo (122, 123, 124). Estos compuestos pueden unirse al ADN y formar aductos que darían como resultado causantes de mutaciones, lo cual aumenta las posibilidades de padecer cáncer (125). Existen estudios realizados sobre los alimentos que analizaron, por ejemplo, el consumo de pescados provenientes de criaderos. Si bien este alimento es considerado una fuente principal de nutrientes, podría presentar dada su procedencia niveles de toxicidad, debido a la influencia de la contaminación ambiental. También está la alerta sobre el consumo de pescados de gran porte que poseen bioacumulación de compuestos orgánicos en su carne. Con las frutas y verduras y consumo de agua ocurre algo similar. La recomendación para aquellos que quieren evitar la exposición a estos químicos y tóxicos es que consuman frutas y verduras orgánicas, de origen certificado, y evitar el consumo de carnes, utilizando barreras aislantes en aquellas actividades laborales en donde puedan verse expuestos (126, 127).

También el azúcar y los edulcorantes son ampliamente consumidos y no existe un consenso sobre la genotoxicidad o el vínculo de estos compuestos con el cáncer. En cualquier caso, vale destacar algunos como el aspartamo, acesulfamo-K que han sido señalados como genotóxicos (128).

Un factor interno muy importante y que explica algunas de las variables encontradas son los radicales libres, especies químicas que poseen uno o más electrones desapareados, capaces de existir en forma independiente y que se producen en todas las células. En la actividad física de alto rendimiento, la sobreproducción de radicales libres por las mitocondrias puede generar una baja de los niveles de las defensas antioxidantes existentes en el organismo humano y ello puede conllevar daño de los tejidos debido al desbalance entre las defensas antioxidantes y los radicales libres, lo que genera, finalmente, estrés oxidativo (129, 130, 131).

Otro factor es el ejercicio físico, que provee beneficios desde lo corporal y, también, a la psiquis. Entre estos encontramos: a) control de peso, b) reducción del riesgo de padecer enfermedades y afecciones (diabetes, presión arterial, entre otros), c) promueve el vínculo social, d) a nivel del medio interno: ayuda al estado anímico, al ciclo de sueño-vigilia (ritmo circadiano), entre otros. Como factor predisponente es aplicable en el ejercicio de alto rendimiento, que trae posibles secuelas (en el largo plazo) sobre el estado de salud del individuo debido a ser un inductor del estrés oxidativo, que se define como una situación de desequilibrio frente a un aumento de especies reactivas de oxígeno o una disminución de antioxidantes (132).

En cuanto a la exposición solar, los rayos UVB colaboran con a) el ritmo circadiano, b) mejoran el rendimiento cognitivo, c) reducen la irritabilidad y la fatiga crónica, d) contribuyen con la producción de la vitamina D, entre otros. Como factor predisponente, Merrill y Frutos (133) explican en su publicación la asociación entre la exposición solar (los rayos UVB) y la producción de la vitamina D con el linfoma de Hodgkin. Ellos aseveran que la exposición a los rayos ultravioleta B permite que el sistema tegumentario sintetice mayor cantidad de vitamina D. Las alteraciones de la producción de esta vitamina están dadas por el rango etario avanzado de los sujetos de la muestra y la raza, ya que, a mayor pigmentación, menor capacidad para producir la vitamina. La evidencia existente entre la exposición solar con el linfoma de Hodgkin demuestra que la vitamina D inducida por los rayos UVB no es un factor protector.

Entre los factores biológicos, se ha demostrado que las hormonas estrógeno y progesterona, sintetizadas por el organismo, estimulan la formación y la proliferación de tumores cancerígenos (p. ej., cánceres que expresan receptores para estas hormonas: el cáncer de seno). Los anticonceptivos son análogos sintéticos de estas hormonas y también podrían aumentar el riesgo de cáncer, como el de cuello uterino (cáncer cervical), al cambiar la susceptibilidad de las células de este a causa de una infección persistente por los tipos de HPV (virus del papiloma humano). Como protectores pueden a) mejorar afecciones de la piel (acné), b) disminuir el riesgo de padecer cáncer de ovario y de endometrio, c) disminuir el dolor menstrual y regular los ciclos ováricos y menstruales, disminuir la probabilidad de embarazo extrauterino, d) ofrecer un efecto protector

a los riesgos de algunos cánceres, como la supresión de la proliferación de las células del endometrio (cáncer de endometrio); reducción del número de ovulaciones que experimenta una mujer en su vida, lo cual reduciría la síntesis de las hormonas estrógeno y progesterona (cáncer de ovario) (134).

Finalmente, Gallagher (135) concluye que los factores predisponentes de cáncer se encuentran más vinculados a una mayor exposición a drogas, químicos, radiación o infecciones; afirmación con la que coincido, puesto que se demostró a partir de los resultados de la correlación negativa que pueden ser factores predisponentes a la enfermedad neoplásica, aunque se tenga el preconcepto de que en realidad son protectores.

Se observa en la evolución de la inteligencia artificial que es una tecnología capaz de simular procesos de la inteligencia del ser humano, tales como el aprendizaje, la autocorrección y el razonamiento gracias a que los algoritmos funcionan emulando las redes neuronales del cerebro humano. Esto resulta clave al aplicar esta tecnología en el ámbito de la salud, de modo que se contribuye en el avance de la medicina, lo que beneficia tanto a los pacientes como a todos los profesionales de la salud.

Desde sus inicios, la aplicación experimental de la IA permitió analizar la información sanitaria como herramienta para gestionar los datos de los pacientes. Muchos de los avances en diagnóstico y pronóstico de enfermedades, como por ejemplo, las infecciosas, renales y reumatológicas, se lograron gracias al procesamiento e interpretación de imágenes por medio de la IA como, así también, la predicción de posibles reacciones adversas o bien la buena adhesión de un paciente al tratamiento propuesto. Estos avances también permiten dar un paso adelante y conocer

los factores predisponentes para desarrollar medidas de prevención con el consiguiente beneficio de un diagnóstico precoz.

Con los resultados de los entrenamientos por aprendizaje supervisado y no supervisado se corroboraron los datos ya conocidos que validaron estas herramientas sobre su eficacia y eficiencia. Dentro de estos análisis, el aporte significativo obtenido por aprendizaje no supervisado fue el de la correlación negativa al demostrar que aquellas medidas que se supone son protectoras en realidad podrían convertirse en predisponentes, lo que revela la importancia de los contaminantes orgánicos sobre los alimentos, mayor exposición a drogas, químicos, radiación o infecciones.

Otros de los beneficios del aprendizaje automático son: facilitar las decisiones clínicas de los profesionales de la salud, contribuir de manera significativa a la prevención y a la detección temprana de enfermedades; y durante el proceso de formación de los profesionales, establecer nuevas formas de estudiar los datos obtenidos (para los que la lógica no tiene alcance), generar regulaciones para el uso de la inteligencia artificial en la medicina, reducir costos y, por sobre todo, elevar la calidad de la atención sanitaria y, en consecuencia, mejorar la calidad de vida de los usuarios.

Con este trabajo se buscaron nuevos factores predisponentes de enfermedad neoplásica. Se considera que todos los modelos desarrollados podrían resultar una herramienta potencialmente útil que permitiría obtener la ventaja del manejo de un gran número de variables y la disminución del tiempo de análisis.

Es primordial remarcar que para este trabajo se usó una base de datos propia, inclusive en la región, lo cual nos permite marcar un precedente.

Bibliografía

1. Lanzagorta-Ortega, D.; Carrillo-Pérez, D. L. & Carrillo-Esper, R. Inteligencia artificial en medicina: presente y futuro. *Gac Med Mex*. 2023 [consultado en agosto de 2023]; 158 (91): 17-21. Disponible en http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0016-38132022001100017&lng=es.
2. Ávila-Tomas, J. F.; Mayer-Pujadas, M. A. & Quesada-Varela, V. J. La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas. *Aten Primaria*. 2021 [consultado en abril de 2023]; 53 (1): 81-8. Disponible en https://dx.doi.org/10.1016/j.aprim.2020.04.014.
3. Luna, M. & Caterine, P. *Detección automática de melanoma aplicando métodos de aprendizaje profundo para el procesamiento y análisis de imágenes dermatológicas*. Programa de ingeniería biomédica; 2021 [consultado en abril de 2023]. Disponible en https://repositorio.escuelaing.edu.co/handle/001/1935.
4. Prado Vázquez, G. *Caracterización molecular del melanoma mediante inteligencia artificial: Implicaciones en pacientes tratados con inmunoterapia*. Universidad Autónoma de Madrid; 2021 [consultado en abril de 2023]. Disponible en http://hdl.handle.net/10486/696364.
5. Meza-Junco, J.; Montaño-Loza, A. & Aguayo-González, Á. Bases moleculares del cáncer. *Rev Invest Clín*. 2006 [consultado en abril de 2023]; 58 (1): 56-70. Disponible en https://www.medigraphic.com/pdfs/revinvcli/nn%2D2006/nn061h.pdf.

6. Soca, M. P. E.; Almaguer Herrera, A.; Ponce de León, D.; Sales Márquez, H. & Pérez Rodríguez, H. El cáncer: una enfermedad genética. *CCM*. 2007 [consultado en agosto de 2023]; 58 (1): 56-70. Disponible en http://www.cocmed.sld.cu/no113/n113rev1.htm.
7. Bermúdez Garcell, A. J.; Serrano Gámez, N. B.; Teruel Ginés, R.; Leyva Montero, M. de los Á. & Naranjo Coronel, A. A. Biología del cáncer. *CCH Correo cient Holguín*. 2019 [consultado en noviembre de 2023]; 23 (4): 1394-1416. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1560-43812019000401394&lng=es.
8. Peralta-Zaragoza, O.; Bahena-Román, M.; Díaz Benítez, C. E.; & Madrid-Marina, V. Regulación del ciclo celular y desarrollo de cáncer: perspectivas terapéuticas. *Rev Salud Pública Mex*. 1997 [consultado en octubre de 2023]; 58 (1): 451-462. Disponible en https://www.scielosp.org/article/spm/1997.v39n5/451-462/#.
9. Aiassa, D. & Bosch, B. Toxicología genética y salud ambiental. 1a ed. Córdoba: CEPYD; Secretaría de Ciencia y Tecnología de la provincia de Córdoba, Argentina; 2015 [consultado en septiembre de 2023]. Disponible en https://www.cba.gov.ar/wp-content/4p96humuzp/2016/05/PROTRI-2013-Libro-Final-AIASSA.pdf.
10. Adrover, C.; Alfaro Galan, L.; Arumi de Dios, M. *et al*. *Factores pronósticos y predictivos del cáncer*. Cap. 1. Sociedad Española de Senología y Patología Mamaria. España. 2017 [consultado en noviembre de 2023]. Disponible en https://es.scribd.com/document/441767664/Factores-pronosticos-y-predictivos-de-cancer-de-mama.

11. Instituto Nacional del Cáncer. *Sustancias en el ambiente que causan cáncer*. 2015 [consultado en noviembre de 2023]. Disponible en https://www.cancer.gov/espanol/cancer/causas-prevencion/riesgo/sustancias.
12. Civetta, M. T.; & Civetta, J. D. *Salud pública Méx*. 2011 [consultado en octubre de 2023]; 53 (5): 405-414. Disponible en http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0036-36342011000500008&lng=es.
13. Organización Mundial de la Salud. *Agents Classifed by the IARC Monographs*, Volumes 1-133. 2023 [consultado en noviembre de 2023]. Disponible en https://monographs.iarc.who.int/list-of-classifications/.
14. National Toxicology Program. Substances listed in the Fifteenth Report on Carcinogens. *Report on Carcinogens, Fifteenth Edition*. [consultado en noviembre de 2023]. Disponible en https://ntp.niehs.nih.gov/ntp/roc/content/listed_substances_508.pdf.
15. Pozzobon, S.; Rivera Nieto, C.; Lasso Ruales, X. T.; Alcaraz Martínez, A. B. & Utria Velázquez A. Relación entre la exposición ocupacional a campos electromagnéticos y riesgo de desarrollar cáncer: una revisión sistemática. *Med. segur. trab*. 2019 [consultado en octubre de 2023]; 65 (255): 139-159. Disponible en http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0465-546X2019000200006&lng=es.
16. Szmigielski, S. & Sobiczewska, E. Risk of neoplastic diseases in conditions of exposure to power magnetic fields – epidemiologic investigations. *Med Pr*. 2009 [consultado en noviembre de 2023]; vol. 60 (3): 223-233. Disponible en https://pubmed.ncbi.nlm.nih.gov/19746891/.

17. Dorado Dellmans, P. J. *Centro de seguridad nuclear*. España. 2010 [consultado en diciembre de 2022]. Disponible en https://www.csn.es/documents/10182/914805/Dosis%20de%20radiaci%C3%B3n.
18. Vivallo, L.; Villanueva, L.; & Sanhueza S. *Efectos de las radiaciones ionizantes en el ser humano. Informe*. Chile: Comisión Chilena de Energía Nuclear, Departamento de Protección Radiológica y Ambiental; 2010. Disponible en https://www.yumpu.com/es/document/view/13079522/efectos-de-las-radiaciones-ionizantes-en-el-ser-humano-comision-.
19. Organización Mundial de la Salud. Radiaciones ionizantes: efectos en la salud y medidas de protección. 2023 [consultado en marzo de 2023]. Disponible en https://www.who.int/es/news-room/fact-sheets/detail/ionizing-radiation-health-effects-and-protective-measures.
20. Bird, W. A. & Braxton Little, J. A tale of two forests: addressing postnuclear radiation at Chernobyl and Fukushima. *Environ Health Perspect*. 2013 [consultado en noviembre de 2023]; 121 (3): 78-85. Disponible en http://dx.doi.org:10.1289/ehp.121-a78.
21. Lesyuk, O.; Sousa, P. E.; Rodrigues, S. I.; Abrantes, A. F.; de Almeida, R.; Pinheiro, J. P. *et al*. Study of scattered radiation during fluoroscopy in hip surgery. *Radiol Bras*. 2016 [consultado en noviembre de 2023]; 49 (4), 234-240. Disponible en https://doi.org/10.1590/0100-3984.2014.0146.
22. Yu, E. & Khan, S. N. Does less invasive spine surgery result in increased radiation exposure? A systematic review. *Clin Orthop Relat Res*. 2014 [consultado en noviembre de 2023]; 472 (6): 1738-1748. Disponible en https://doi.org/10.1007/s11999-014-3503-3.

23. Mother To Baby | Fact Sheets. Brentwood (TN): Organization of Teratology Information Specialists (OTIS). Radiación ionizante en entornos de trabajo. 1994 [consultado en diciembre de 2022]. Disponible en https://www.ncbi.nlm.nih.gov/books/NBK582774/.
24. Gaytán Fernández, S.; Barragán-Hervella, R. G.; Quiroz-Williams, J.; Rodríguez Palacios, C. L. & Sánchez-González, G. Exposición a radiación ionizante en médicos residentes de ortopedia en un hospital de referencia. *Cir Cir*. 2023 [consultado en diciembre de 2022]; 91 (1): 64-72. Disponible en http://dx.doi.org/10.24875/ciru.21000644.
25. Garnacho Saucedo, G. M.; Salido Vallejo, R. & Moreno Giménez, J. C. Effects of solar radiation and an update on photoprotection. *An Pediatr (Engl Ed)*. 2020 [consultado en noviembre de 2023]; 92 (6): 377. Disponible en https://www.sciencedirect.com/science/article/pii/S2341287920300752.
26. Heistein, J. B.; Acharya, U.; & Mukkamalla, S. K. R. Malignant Melanoma. *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2023 [consultado en octubre de 2023]. Disponible en https://www.ncbi.nlm.nih.gov/books/NBK470409/.
27. Lewandowska, A. M.; Rudzki, M.; Rudzki, S.; Lewandowski, T. & Laskowska, B. Environmental risk factors for cancer – review paper. *Ann Agric Environ Med*. 2019 [consultado en diciembre de 2022]; 26 (1): 1-7. Disponible en https://doi.org/10.26444/aaem/94299.
28. Shingala, U. Environmental factors and terminal cancer development – review article. *Int J of Innov Sci and Res Technol*. 2020 [consultado en diciembre de 2023]; 5 (7). Disponible en https://ijisrt.com/assets/upload/files/IJISRT20JUL859.pdf.

29. Philippe, A. & Doré, J. F. Ultraviolet radiation and cutaneous melanoma: a historical perspective. *Melanoma Research*. 2020 [consultado en septiembre de 2022]; 30 (2): 113-125. Disponible en https://dx.doi.org/10.1097/cmr.0000000000000609.
30. Matta, J. L.; Ramos, J. M.; Armstrong, R. A. & D'Antoni, H. Niveles de luz ultravioleta ambiental asociados con apoptosis y necrosis en fibroblastos humanos. *Acta Bioquím Clín Latinoam*. 2006 [consultado en septiembre de 2023]; 40 (4): 453-460. Disponible en http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325-29572006000400015&lng=es.
31. Moan, J.; Porojnicu, A. C. & Dahlback, A. Ultraviolet radiation and Malignant Melanoma. En *Sunlight, vitamin D and skin cancer*. New York, NY: Springer New York. 2009; vol. 624: 104-116. Disponible en https://doi.org/10.1007/978-0-387-77574-6_9.
32. Berwick, M.; Pestak, C. & Thomas, N. Solar ultraviolet exposure and mortality from skin tumors. En *Sunlight, Vitamin D and Skin Cancer*. New York, NY: Springer New York. 2014; vol. 810: 342-358. https://doi.org/10.1007/978-1-4939-0437-2_19.
33. De Gruijl, F. R. Photocarcinogenesis: UVA vs UVB. En *Methods in enzymology*. Elsevier. 2000; vol. 319: 359-366. Disponible en https://doi.org/10.1016/s0076-6879(00)19035-4.
34. Schwartz, R. Cáncer de piel no melanoma en cabeza y cuello. *Rev Méd Clín Las Condes*. 2018 [consultado en enero de 2023]; 29 (4): 455-467. Disponible en https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-cancer-piel-no-melanoma-cabeza-S0716864018300889.

35. Estadísticas – Incidencia. Argentina.gob.ar. 2019 [consultado en enero de 2023]. Disponible en https://www.argentina.gob.ar/salud/instituto-nacional-del-cancer/estadisticas/incidencia.
36. Seitz, H. K. & Becker, P. Alcohol metabolism and cancer risk. *Alcohol Res Health*. 2007; 30 (1): 38-41, 44-47. Disponible en https://pubmed.ncbi.nlm.nih.gov/17718399/.
37. Shield, K. D.; Rylett, M.; Gmel, G.; Kehoe-Chan, T. A. & Rehm, J. Global alcohol exposure estimates by country, territory and region for 2005: a contribution to the Comparative Risk Assessment for the 2010 Global Burden of Disease Study. Addiction. 2013; 108: 912-922. [consultado en agosto de 2023]. Disponible en https://pubmed.ncbi.nlm.nih.gov/23347092/.
38. Veitía, G.; Otero, J.; Ruiz, N.; Graterol, A. & Brizuela, L. Consumo de tabaco y alcohol en pacientes con cáncer esofágico y su relación con el tipo histológico: Hospital Vargas de Caracas período 2004-2009. G E N. 2010 [citado el 30 de enero de 2023]; 64 (1): 37-39. Disponible en http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0016-35032010000100010&lng=es.
39. Varela Centelles, P. I.; Romero Méndez, A.; Varela Centelles, A. & Rodríguez González, V. Riesgo de cáncer oral atribuible al consumo de alcohol y tabaco en la Zona de Salud de Burela (Lugo). *RCOE*. 2007 [consultado el 30 de enero de 2023]; 12 (3): 177-181. Disponible en http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1138-123X2007000200004&lng=es.
40. Grupo de trabajo de la IARC sobre la evaluación de riesgos cancerígenos para los seres humanos. *Tabaco sin humo y algunas N-nitrosaminas específicas*

del tabaco. Oxford, Inglaterra: Agencia Internacional para la Investigación del Cáncer. 2007. Disponible en https://www.ncbi.nlm.nih.gov/books/NBK326497/.

41. Thomas, J. L. *et al*. Metabolites of a tobacco-specific lung carcinogen in children exposed to second-hand or thirdhand tobacco smoke in their homes. *Cancer Epidemiol Biomarkers Prev*. 2011; 20 (6): 1213-1221. Disponible en https://doi.org/10.1158/1055-9965.EPI-10-1027.
42. Jung, S. Y.; Yu, H.; Pellegrini, M.; Papp, J. C.; Sobel, E. M. & Zhang, Z.-F. Genetically determined elevated C-reactive protein associated with primary colorectal cancer risk: Mendelian randomization with lifestyle interactions. *American Journal of Cancer Research*. 2021; 11 (4): 1733-1753 [consultado en julio de 2023]. Disponible en https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8085861/.
43. Chen, X.; Jansen, L.; Guo, F.; Hoffmeister, M.; Chang-Claude, J. & Brenner H. Smoking, genetic predisposition, and colorectal cancer risk. *Clin Transl Gastroenterol*. 2021; 12 (3): e00317. Disponible en https://doi.org/10.14309/ctg.0000000000000317.
44. Rumgay, H.; Shield, K.; Charvat, H.; Ferrari, P.; Sornpaisarn, B. & Obot, I. *et al*. Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. *Lancet Oncol*. Publicado en línea el 14 de julio de 2021. Disponible en https://doi.org/10.1016/S1470-2045(21)00279-5.
45. El alcohol y el cáncer. Cdc.gov. 2023 [consultado el 21 de enero de 2023]. Disponible en https://www.cdc.gov/spanish/cancer/alcohol/index.htm.

46. Becher, H.; Belau, M.; Winkler, V. & Aigner, A. Estimating lung cancer mortality attributable to second hand smoke exposure in Germany. *Int J Public Health*. 2018; 63 (3): 367-375. Disponible en https://doi.org/10.1007/s00038-017-1022-1.
47. Global Burden of Disease [base de datos]. Washington, DC: Institute of Health Metrics. 2019 [consultado el 17 de febrero de 2023]. Disponible en https://extranet.who.int/ncdsmicrodata/index.php/catalog/270.
48. Price, L. R. & Martinez, J. Cardiovascular, carcinogenic and reproductive effects of nicotine exposure: A narrative review of the scientific literature. *F1000Res*. 2020; 8: 1586. Disponible en https://doi.org/10.12688/f1000research.20062.2.
49. IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. *IARC Monogr Eval Carcinog Risks Hum*. 2010; 94. Disponible en https://pubmed.ncbi.nlm.nih.gov/21141240/.
50. Zhang, J.; Bai, R.; Li, M.; Ye, H.; Wu, C.; Wang, C. *et al*. Excessive miR-25-3p maturation via N6-methyladenosine stimulated by cigarette smoke promotes pancreatic cancer progression. *Nat Commun*. 2019; 10 (1): artículo 1858. Disponible en https://doi.org/10.1038/s41467-019-09712-x.
51. Wacławek, S.; Silvestri, D.; Hrabák, P.; Padil, V. V. T.; Torres-Mendieta, R.; Wacławek, M. *et al*. Chemical oxidation and reduction of hexachlorocyclohexanes: A review. *Water Res*. 2019; 162: 302-319. Disponible en http://dx.doi.org/10.1016/j.watres.2019.06.072.
52. Mastandrea, C.; Chichizola, C.; Ludueña, B.; Sánchez, H.; Álvarez, H. & Gutiérrez, A. Hidrocarburos aromáticos policíclicos. Riesgos para la salud y marcadores biológicos. *Acta Bioquim Clin Latinoam*. 2005 [consultado

el 18 de febrero de 2023]; 39 (1): 27-36. Disponible en http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325957200500010006&lng=es.

53. Qué hay en un cigarrillo. ASAT – Asociación Argentina de Tabacología; 2017. Disponible en http://www.asat.org.ar/que-hay-en-un-cigarrillo/.
54. Chen, X.; Li, H.; Mandic, M.; Hoffmeister, M.; & Brenner, H. Assessment of body mass index, polygenic risk score, and development of colorectal cancer. *JAMA Netw Open*. 2022; 5 (12): e2248447. Disponible en http://dx.doi.org/10.1001/jamanetworkopen.2022.48447.
55. Chen, X.; Jansen, L.; Guo, F.; Hoffmeister, M.; Chang-Claude, J. & Brenner, H. Smoking, genetic predisposition, and colorectal cancer risk. *Clin Transl Gastroenterol*. 2021; 12 (3): e00317. Disponible en http://dx.doi.org/10.14309/ctg.0000000000000317.
56. Workowski, K. A.; Bachmann, L. H.; Chan, P. A.; Johnston, C. M.; Muzny, C. A.; Park, I. *et al*. Sexually transmitted infections treatment guidelines, 2021. *MMWR. Recomm Rep*. 2021; 70 (4): 1-187. Disponible en https://doi.org/10.15585/mmwr.rr7004a1.
57. Cornelius, M. E.; Wang, T. W.; Jamal, A.; Loretan, C. G. & Neff, L. J. Tobacco product use among adults – United States, 2019. *MMWR. Morb Mortal Wkly Rep*. 2020; 69 (46), 1736-42. Disponible en https://doi.org/10.15585/mmwr.mm6946a4.
58. Hartwig, A.; Arand, M.; Epe, B.; Guth, S.; Jahnke, G.; Lampen, A. *et al*. Mode of action-based risk assessment of genotoxic carcinogens. *Arch Toxicol*. 2020; 94 (6): 1787-877. Disponible en https://doi.org/10.1007/s00204-020-02733-2.

59. Cogliano, V. J.; Baan, R.; Straif, K.; Grosse, Y.; Lauby-Secretan, B.; El Ghissassi, F. *et al*. Preventable exposures associated with human cancers. *J Natl Cancer Inst*. 2011; 103 (24): 1827-1839. Disponible en https://doi:10.1093/jnci/djr483.
60. Ministry of Health. Dioxins: A technical guide. 2020 (10th edn). Publicado el 20 de septiembre, Ministerio de Salud, Wellington, Nueva Zelanda.
61. Eseknazi, B.; Warner, M.; Brambilla, P.; Signorini, S.; Ames, J. & Mocarelli P. The Seveso accident: A look at 40 years of health research and beyond. *Environ Int*. 2018; 121: 71-84. Disponible en http://dx.doi.org/10.1016/j.envint.2018.08.051.
62. Wahlang, B. Exposure to persistent organic pollutants: impact on women's health. *Rev Environ Health*. 2018; 33 (4): 331-348. Disponible en https://doi.org/10.1515/reveh-2018-0018.
63. Resumen de Salud Pública: Hidrocarburos aromáticos policíclicos (HAP) [Polycyclic Aromatic Hydrocarbons (PHA)]. Cdc.gov. 2021. Disponible en https://www.atsdr.cdc.gov/es/phs/es_phs69.html.
64. Araujo, J.; Yegres, F.; Barreto, G.; Antequera, A. A.; Depool, B. & Rojas, Y. Biocatalizadores fúngicos hidrocarbonoclásticos del género Aspergillus para la descontaminación de agua con Hidrocarburos Policíclicos Aromáticos (HPAs). *Rev cub quím*. 2016; 28 (2): 703-735. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2224-54212016000200013&lng=es&tlng=es.
65. Li, Y. Modern epigenetics methods in biological research. *Methods*. 2021; 187: 104-113. Disponible en http://dx.doi.org/10.1016/j.ymeth.2020.06.022.

66. Agudo, A.; Aronson, K. J.; Bonefeld-Jorgensen, E. C.; Cocco, P.; Cogliana, V. & Cravedi, J. P. International Agency for Research on Cancer. Monographs on the evaluation of carcinogenic risks to humans. Polychlorinated biphenyls and polybrominated biphenyls. A review of human carcinogens. Part F. World Health Organization. Lyon, France: International Agency for Research on Cancer; 2016. Disponible en https://cah.au.dk/publications.
67. Vives, I.; Grimalt, J. O. & Guitart, R. Los hidrocarburos aromáticos policíclicos y la salud humana. *Apuntes de Ciencia y Tecnología*. 2001; 3. Disponible en https://spencervivequimica.jimdofree.com/app/download/8398823369/HIDROCARBUROS+AROM%C3%81TICOS+Y+LA+SALUD.pdf?t=1378134966.
68. Huang, R.; Yuan, X.; Yan, L.; Han, L.; Bao, W.; Chang, L. *et al*. Carbon precursors in coal tar: Extraction and preparation of carbon materials. *Sci Total Environ*. 2021; 788: 147697. Disponible en https://doi.org/10.1016/j.scitotenv.2021.147697.
69. Borda, L. J.; Perper, M. & Keri, J. E. Treatment of seborrheic dermatitis: a comprehensive review. *J Dermatol Treat*. 2019; 30 (2): 158-69. Disponible en https://doi.org/10.1080/09546634.2018.1473554.
70. Sekhon, S.; Jeon, C.; Nakamura, M.; Afifi, L.; Yan, D.; Wu, J. J. *et al*. Review of the mechanism of action of coal tar in psoriasis. *J Dermatolog Treat*. 2018; 29 (3): 230-232. Disponible en https://doi.org/10.1080/09546634.2017.1369494.
71. Whysner, J. Benzene-induced genotoxicity. *J Toxicol Environ Health, Part A*. 2000; 61 (5-6): 347-351. Disponible en https://doi.org/10.1080/00984100050166343.

72. Cruz Campos, F. J. & Delgado Jacobo, D. P. Obesidad como factor de riesgo para cáncer. *Psic*. 2019; 9 (33). Disponible en https://doi.org/10.22201/fesz.20075502e.2019.9.33.80692.
73. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Benzene. Lyon: International Agency for Research on Cancer; 2018. Disponible en https://pubmed.ncbi.nlm.nih.gov/31769947/.
74. Guo, H.; Ahn, S. & Zhang, L. Benzene-associated immunosuppression and chronic inflammation in humans: a systematic review. *Occup Environ Med*. 2021; 78 (5): 377-384. Disponible en http://dx.doi.org/10.1136/oemed-2020-106517.
75. Zubeldia Lauzurica, L. & Gomar Fayos, J. Acrilamida en patatas fritas y productos de aperitivo elaborados en la Comunidad Valenciana. *Gac Sanit*. 2007; 21 (4):334-337. Disponible en https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0213-91112007000400012.
76. Eyre, H.; Kahn, R. & Robertson, R. M., on behalf of the ACS/ADA/AHA Collaborative Writing Committee. Preventing cancer, cardiovascular disease, and diabetes. *Diabetes Care*. 2004; 27 (7): 1812-1824. Disponible en https://pubmed.ncbi.nlm.nih.gov/15220271/.
77. Lajous, M.; Lazcano-Ponce, E.; Hernandez-Avila, M.; Willett, W. & Romieu, I. Folate, vitamin B6, and vitamin B12 intake and the risk of breast cancer among Mexican women. *Cancer Epidemiol Biomarkers Prev*. 2006; 15 (3): 443-448. Disponible en https://doi.org/10.1158/1055-9965.EPI-05-0532.
78. Fung, T. T.; Hu, F. B.; Hankinson, S. E.; Willett, W. C. & Holmes, M. D. Low-carbohydrate diets, dietary approaches to stop hypertension-style diets, and the

risk of postmenopausal breast cancer. *Am J Epidemiol*. 2011; 174 (6): 652-660. Disponible en https://doi.org/10.1093/aje/kwr148.

79. Farvid, M. S.; Cho, E.; Chen, W. Y.; Eliassen, A. H. & Willett, W. C. Adolescent meat intake and breast cancer risk. *Int J Cancer*. 2015; 136 (8): 1909-1920. Disponible en https://doi.org/10.1002/ijc.29218.
80. Lerda, D. E. La acrilamida en los alimentos y la salud humana. Revisión. *Methodo Investig Apl Las Cienc Biol*. 2020; 5 (3): 100-107. Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=7537941.
81. Chaves Ullate, C.; Irias Mata, A. & Arias Echandi, M. L. Formación de acrilamida durante el procesamiento de alimentos. Una revisión. *Rev costarric salud pública*. 2016; 25 (2): 28-35. Disponible en http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-14292016000200028&lng=en.
82. US Department of Agriculture and US Department of Health and Human Services. Dietary Guidelines for Americans, 2020-2025. 9th Ed. [consultado en julio de 2023]. Disponible en www.dietaryguidelines.gov/sites/default/files/2020-12/Dietary_Guidelines_for_Americans_2020-2025.pdf.
83. Gamero, A. & del Carmen, M. La tributación de las bebidas azucaradas como medida para reducir la obesidad en España. Análisis y valoración de la eficacia a partir de su aplicación en otros países. *Rev Bioet Derecho*. 2018 [consultado el 27 de marzo de 2023]; (42): 269-310. Disponible en http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1886-58872018000100015&lng=es.

84. Martínez-Velarde, D.; Málaga-Chávez, R.; & Bernabe-Ortiz, A. Consumo de bebidas azucaradas, verduras y frutas en sujetos con alteración del metabolismo de la glucosa. *Rev Esp Nutr Hum Diet*. 2021 [consultado el 27 de marzo de 2023]; 25(3): 326-336. Disponible en https://dx.doi.org/10.14306/renhyd.25.3.1258.
85. Llaha, F.; Gil-Lespinard, M.; Unal, P.; de Villasante, I.; Castañeda, J.; & Zamora-Ros, R. Consumption of sweet beverages and cancer risk. A systematic review and meta-analysis of observational studies. *Nutrients*. 2021; 13 (2): 516. Disponible en http://dx.doi.org/10.3390/nu13020516.
86. Yu, F.; Jin, Z.; Jiang, H.; Xiang, C.;Tang, J.; Li, T. *et al*. Tea consumption and the risk of five major cancers: a dose-response meta-analysis of prospective studies. *BMC Cancer*. 2014; 14 (1). Disponible en http://dx.doi.org/10.1186/1471-2407-14-197.
87. Wang, Y.; Chen, J.; Zhao, R.; Xia, L.; Cui, Y.-P.; Rao, Z.-Y. *et al*. Dose-response meta-analysis of coffee consumption and risk of colorectal adenoma. *Eur J Clin Nutr*. 2020; 74 (2): 297-306. Disponible en https://pubmed.ncbi.nlm.nih.gov/31308476/.
88. Kohler, L. N.; Garcia, D. O.; Harris, R. B.; Oren, E.; Roe, D. J. & Jacobs, E.T. Adherence to diet and physical activity cancer prevention guidelines and cancer outcomes: A systematic review. *Cancer Epidemiol Biomarkers Prev*. 2016; 25 (7): 1018-1028. Disponible en http://dx.doi.org/10.1158/1055-9965.epi-16-0121.
89. Caballero Torres, Á. E. *et al*. *Temas de higiene de los alimentos*. La Habana: Editorial Ciencias Médicas; 2008. Disponible en https://sceqa.files.wordpress.com/2012/05/libro-higiene-de-alimentos.pdf.

90. Manfredi, E. A. & Rivas, M. Brote de intoxicación alimentaria en un jardín de infantes de la provincia de Buenos Aires. *Rev Argent Microbiol*. 2019; 51 (4): 354-358. Disponible en http://dx.doi.org/10.1016/j.ram.2018.08.008.
91. Aesan – Agencia Española de Seguridad Alimentaria y Nutrición. Gob.es. [consultado en abril de 2023]. Disponible en https://www.aesan.gob.es/AECOSAN/web/seguridad_alimentaria/subdetalle/micotoxinas.htm.
92. Mctiernan, A.; Friedenreich, C. M.; Katzmarzyk, P. T.; Powell, K. E.; Macko, R. & Buchner, D. *et al*. Physical activity in cancer prevention and survival: A systematic review. *Med Sci Sports Exerc*. 2019; 51 (6): 1252-1261. Disponible en http://dx.doi.org/10.1249/mss.0000000000001937.
93. Martínez Miranda, M. M.; Vargas del Río, L. M. & Gómez Quintero, V. M. Aflatoxinas: incidencia, impactos en la salud, control y prevención. *Biosalud*. 2013 [consultado en abril de 2023]; 12 (2): 89-109. Disponible en http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-95502013000200008&lng=en.
94. Rojas Jaimes, J.; Chacón-Cruzado, M.; Díaz-Tello, A. & Castañeda-Pelaez, L. Quantification of carcinogenic aflatoxins in unprocessed foods and their implication for consumption in Lima, Peru. *Nutr Hosp*. 2020 [consultado en mayo de 2023]. Disponible en http://dx.doi.org/10.20960/nh.03240.
95. Bolet Astoviza, M. & Socarrás Suárez, M. Micotoxinas y cáncer. *Rev Cuba Investig Bioméd*. 2005 [consultado en abril de 2023]; 24 (1): 54-59. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-03002005000100007&lng=es.

96. Perusia, O. R. & Rodríguez, A. R. Micotoxicosis. *Rev Investig Vet Peru*. 2001 [consultado en septiembre de 2023]; 12 (2): 87-116. Disponible en http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1609-91172001000200013&lng=es.
97. Vega, D. R. P. V. Toxicología de alimentos. *Edu.ar* [consultado en mayo de 2023]. Disponible en https://www.fio.unicen.edu.ar/usuario/gmanrique/images/Toxicologia_de_Alimentos_VegaFlorentino.pdf.
98. Szkaradkiewicz, A. A. Microbes and oncogenesis. *Contemp Oncol (Pozn)*. 2003; 7 (2): 96-101. Disponible en https://www.termedia.pl/Microbes-and-oncogenesis,3,297,0,1.html.
99. González, C. A. & Agudo, A. Carcinogenesis, prevention and early detection of gastric cancer: Where we are and where we should go. *Int J Cancer*. 2012; 130: 745-753. Disponible en http://dx.doi.org/10.1002/ijc.26430.
100. Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A. *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021; 71 (3): 209-249. Disponible en http://dx.doi.org/10.3322/caac.21660.
101. Cameán, A. M. & Repetto, M. Toxicología alimentaria. Pdfcoffee.com. 2019. [citado el 6 de marzo de 2023]. Disponible en https://pdfcoffee.com/camean-ana-m-y-repetto-manuel-toxicologia-alimentaria-pdf-free.html.
102. Ángel, G. *Tratado de nutrición (tomo 2). Composición y calidad nutritiva de los alimentos* (2ª ed.). Casassa y Lorenzo. Disponible en https://www.casassaylorenzo.com/9788498353471/TRATADO+DE+NUTRI-

CION+(TOMO+2)+COMPOSICION+Y+CALIDAD+NUTRITIVA+DE+LOS+ALIMENTOS+(2+EDICION).

103. McGuire, S. World cancer report 2014. Geneva, Switzerland: World Health Organization, International Agency for Research on Cancer, WHO press, 2015. *Adv Nutr*. 2016; 7 (2): 418-419. Disponible en http://dx.doi.org/10.3945/an.116.012211.
104. EFSA Panel on Contaminants in the Food Chain (CONTAM). Scientific Opinion on acrylamide in food. *EFSA J*. 2015; 13 (6): 1-49. Disponible en https://cot.food.gov.uk/sites/default/files/2022-12/TOX-2022-66%20Acrylamde%20Acc%20V%20.pdf.
105. Zhang, J.; Bai, R.; Li, M.; Ye, H.; Wu, C., Wang, C, *et al*. Excessive miR-25-3p maturation via N6-methyladenosine stimulated by cigarette smoke promotes pancreatic cancer progression. *Nat Commun*. 2019; 10 (1): 1-14. Disponible en http://dx.doi.org/10.1038/s41467-019-09712-x.
106. Pietropaoli, F.; Pantalone, S.; Cichelli, A.; & d'Alessandro, N. Acrylamide in widely consumed foods – a review. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*. 2022 [consultado el 18 de septiembre de 2023]; 39 (5): 853-887. Disponible en https://pubmed.ncbi.nlm.nih.gov/35286246/.
107. Michalak, M.; Czarnowska-Kujawska, M.; Klepacka, J. & Gujska, E. Effect of microwave heating on the acrylamide formation in foods. *Molecules*. 2020 [citado el 12 de septiembre de 2023]; 25(18):4140. Disponible en https://www.mdpi.com/1420-3049/25/18/4140.

108. Jakszyn, P.; Agudo, A.; Berenguer, A.; Ibáñez, R.; Amiano, P.; Pera, G. *et al*. Intake and food sources of nitrites and N-nitrosodimethylamine in Spain. *Public Health Nutr*. 2006; 9: 785-789 1. Disponible en https://doi.org/10.1079/phn2005884.
109. Li, X.; Bei, E.; Qiu, Y.; Xiao, H.; Wang, J.; Lin, P. *et al*. Intake of volatile nitrosamines by Chinese residents in different provinces via food and drinking water. *Sci Total Environ*. 2021; 754: 142121. Disponible en http://dx.doi.org/10.1016/j.scitotenv.2020.142121.
110. Duarte, L. E.; Delgado, F.; Di Leo, N. C.; Bertone, C. L.; Franci Alvarez, M.; Montico, S. *et al*. Mortalidad por cáncer, arsénico y nitratos en aguas de consumo y superficies sembradas en Argentina. *Rev Panam Salud Pública*. 2022; 46: 1. Disponible en http://dx.doi.org/10.26633/rpsp.2022.129.
111. Global Burden of Disease (GBD). The Institute for Health Metrics and Evaluation. [consultado el 13 de mayo de 2023]. Disponible en https://www.healthdata.org/research-analysis/gbd.
112. Barrionuevo Cornejo, C. & Dueñas Hancco, D. Clasificación actual del carcinoma de pulmón: consideraciones histológicas, inmunofenotípicas, moleculares y clínicas. *Horiz Med*. 2019; 19 (4): 74-83. Disponible en http://dx.doi.org/10.24265/horizmed.2019.v19n4.11.
113. De Groot, P. M.; Wu, C. C.; Carter, B. W. & Munden, R. F. The epidemiology of lung cancer. *Transl Lung Cancer Res*. 2018 [consultado el 13 de mayo de 2023]; 7 (3): 220-33. Disponible en https://pubmed.ncbi.nlm.nih.gov/30050761/.
114. Al monte, M.; Murillo, R.; Sánchez, G. I.; González, P.; Ferrera, A.; Picconi, M. A. *et al*. Multicentric study of cervical cancer screening with human papillomavirus testing and assessment of triage methods in

Latin America: the ESTAMPA screening study protocol. *BMJ Open*. 2020; 10 (5): e035796. Disponible en http://dx.doi.org/10.1136/bmjopen-2019-035796.

115. Clasificación de carcinógenos IARC.OMS. *Who.int* [consultado en abril de 2023]. Disponible en https://monographs.iarc.who.int/list-of-classifications/.
116. Inoue-Choi, M.; Sinha, R.; Gierach, G. L. & Ward, M. H. Red and processed meat, nitrite, and heme iron intakes and postmenopausal breast cancer risk in the NIH-AARPDiet and Health Study: Meat Intake and Postmenopausal Breast Cancer. *Int J Cancer*. 2016 [consultado el 2 de abril de 2023]; 138 (7): 1609-1618. Disponible en http://dx.doi.org/10.1002/ijc.29901.
117. Workowski, K. A.; Bachmann, L. H.; Chan, P. A.; Johnston, C. M.; Muzny, C. A.; Park, I. *et al*. Sexually transmitted infections treatment guidelines, 2021. *MMWR Recomm Rep*. 2021; 70 (4): 1-187. Disponible en http://dx.doi.org/10.15585/mmwr.rr7004a.
118. Li, Y. & Hecht, S. S. Carcinogenic components of tobacco and tobacco smoke: A 2022 update. *Food Chem Toxicol*. 2022; 165: 113179. Disponible en http://dx.doi.org/10.1016/j.fct.2022.113179.
119. Teuwen, J. & Moriakov, N. Convolutional neural networks. En *Handbook of Medical Image Computing and Computer Assisted Intervention*. 1st ed. Elsevier. 2020; 481-501.
120. Casanova, R.; Saldana, S.; Chew, E. Y.; Danis, R. P.; Greven, C. M. & Ambrosius, WT. Application of random forests methods to diabetic retinopathy classification analyses. *PLoS One*. 2014; 9 (6): e98587. Disponible en http://dx.doi.org/10.1371/journal.pone.0098587.

121. Criminisi, A. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found Trends® Comput Graph Vis*. 2011; 7 (2-3): 81-227. Disponible en https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/CriminisiForests_Found-Trends_2011.pdf.
122. Fariña-Tojo, J. Cálculo de la entropía producida en diversas zonas de Madrid. *Cuad Investig Urbanística*. Upm.es 1998 [consultado en enero de 2023]. Disponible en https://oa.upm.es/56253/1/Entropia_Madrid_2a_ed.pdf.
123. Hernández, J. A. R.; Herrera, D. M. R. & Rodríguez, J. E. A research comparative among association rules algorithms. *Vis Electron*. 2016 [consultado el 30 de abril de 2023]; 10 (2): 7. Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=6081876.
124. Murphy, N.; Xu, L.; Zervoudakis, A.; Xue, X.; Kabat, G.; Rohan, T. E. *et al*. Reproductive and menstrual factors and colorectal cancer incidence in the Women's Health Initiative Observational Study. *Br J Cancer*. 2017; 116 (1): 117-125. Disponible en https://pubmed.ncbi.nlm.nih.gov/27898658/.
125. Corrales, L. C. & Muñoz Ariza, M. M. Estrés oxidativo: origen, evolución y consecuencias de la toxicidad del oxígeno. *Nova*. 2012; 10 (18): 213-225. Disponible en http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-24702012000200009&lng=en.
126. Fainstein, M. K. *Radicales libres y estrés oxidativo: aplicaciones médicas*. Editorial El Manual Moderno; 2008.
127. Cervellin, G.; Comelli, I.; Benatti, M.; Sanchis-Gomar, F.; Bassi, A.; & Lippi, G. Non-traumatic rhabdomyolysis: Background, laboratory features, and acute clinical

management. *Clin Biochem*. 2017; 50 (12): 656-662. Disponible en https://pubmed.ncbi.nlm.nih.gov/28235546/.

128. Ortiz Escarza, J. M. & Medina López, M. E. Estrés oxidativo, ¿un asesino silencioso? *Educ quím*. 2020; 31 (1): 2. Disponible en https://www.revistas.unam.mx/index.php/req/article/view/69709.
129. Varela-Moreiras, G.; Samaniego Vaesken, M. de L. & Partearroyo, T. Presence and consumption of sugars and low and no-calorie sweeteners in the Spanish diet: an updated overview. *Nutr Hosp*. 2019; 36 (Spec3). Disponible en https://pubmed.ncbi.nlm.nih.gov/31368329/
130. Debras, C.; Chazelas, E.; Srour, B.; Druesne-Pecollo, N.; Esseddik, Y.; Szabo de Edelenyi, F. *et al*. Artificial sweeteners and cancer risk: Results from the NutriNet-Santé population-based cohort study. *PLoS Med*. 2022; 19 (3): e1003950. Disponible en https://pubmed.ncbi.nlm.nih.gov/35324894/.
131. Pavanello, S.; Moretto, A.; La Vecchia, C. & Alicandro, G. Non-sugar sweeteners and cancer: Toxicological and epidemiological evidence. *Regul Toxicol Pharmacol*. 2023; 139: 105369. Disponible en http://dx.doi.org/10.1016/j.yrtph.2023.105369.
132. Cavagnari, B. M.; Gómez, G.; Kovalskys, I.; Quesada, D.; Brenes, J. C. & grupo ELANS. Consumo de edulcorantes no calóricos en la población adulta de Argentina *[Non-caloric sweeteners consumption in the adult population of Argentina]. Medicina (B Aires)*, 82 (6), 881-890. Disponible en https://pubmed.ncbi.nlm.nih.gov/36571527/.

133. Merrill, R. M. & Frutos, A. M. Ecological evidence for lower risk of lymphoma with greater exposure to sunlight and higher altitude. *High Alt Med Biol*. 2020; 21 (1): 37-44. Disponible en https://pubmed.ncbi.nlm.nih.gov/31765244/.
134. Heredia-García, G.; Gómez-Oliván, L. M.; Orozco-Hernández, J. M.; Luja-Mondragón, M.; Islas-Flores, H.; SanJuan-Reyes, N. *et al*. Alterations to DNA, apoptosis and oxidative damage induced by sucralose in blood cells of Cyprinus carpio. *Sci Total Environ*. 2019; 692: 411-421. Disponible en http://dx.doi.org/10.1016/j.scitotenv.2019.07.165.
135. Gallagher, J. C.; Peacock, M.; Yalamanchili, V. & Smith, L. M. Effects of vitamin D supplementation in older African American women. *J Clin Endocrinol Metab*. 2013; 98 (3): 1137-1146. Disponible en https://pubmed.ncbi.nlm.nih.gov/23386641/.

Anexos

Anexo I. Codificación de las variables

Cuantificación de variables nominales

Los algoritmos de IA utilizan datos del tipo numérico como entrada. Sin embargo, muchos datos obtenidos en la encuesta son del tipo nominal. Por ello, para incorporarlos en el estudio, se requiere transformarlos en números.

Una técnica muy común se basa en asignar valor 1 cuando una variable binaria adquiere el valor "Verdadero" y 0 cuando es "Falso". Sin embargo, esta no es la única estrategia.

Aquí se presentan las definiciones o criterios tomados para la cuantificación de algunas variables nominales según el caso. Los criterios fueron consensuados con profesionales idóneos.

I) Horas de ejercicio o actividad física

Con relación a la cantidad de veces, se consideraron las siguientes opciones: nunca, no contesta, 1 vez al mes, entre 2 y 4 veces por mes, una vez por semana, entre una y dos veces por semana, todos los días.

En cada ocasión, se considera la cantidad de horas: nunca, no contesta, 1 h, entre 1 h y 2 h.

Con ello, se presenta una variable cuantitativa continua que se obtiene a partir de la fórmula:

Ejercicios horas por mes = cantidad de veces * cantidad de horas

Teniendo en cuenta que nunca = no contesta = 0 veces por mes, 1 vez al mes equivale a 1/30 veces por mes. Entre 1 y 2 veces = 1,5 veces por mes = 1/20 veces por mes. Entre 2 y 4 veces por mes = 3 veces por mes. 1 vez por semana = 4,35 veces por mes. Entre 1 y 2 veces por semana = 1,5 veces por semana = 6,525 veces por mes. Todos los días = 30 veces por mes. Por otro lado, para la cantidad de horas se tuvo en cuenta: "nunca o no contesta" = 0 horas. "1 hora" = 1 hora, "entre 1 y 2 horas" = 1,5 horas.

II) Uso_de_cama_solar

Valores: "nunca" = 0, "1 vez" = 1, "entre 2 y 4 veces" = 2, "más de 4 veces" = 3.

Cama_solar (en h/ semana): "nunca o no contesta" = 0 h, "1 vez" = 2h, "entre 2 y 4 veces" = 6 h, "más de 4 veces" = 10 h. **Prebronceado** (en h/semana): "no" = 0h, "casi nunca" = 0,5 h, "a veces" = 1h, "casi siempre" = 2h, "siempre" = 3h.

III) Toma_sol_ h_mes

Horario sol: antes de las 11 h = después de las 16 h = 3,5 h. Entre 12 h y 15 h = 7 h.

IV) Usa_bloqueador (coeficiente)

"Nunca o NULL" = 1, "casi nunca" = 0,8, "a veces" = 0,7, "siempre" = 0,5, toma sol h mes = ⅓ * 4,125 * (cama solar + prebronceado + horario sol) * bloqueador.

V) Uso_sahumerios_mes

"Nunca o no contesta" = 0, "entre 1 y 3 veces por semana" = 9, "más de 3 veces por semana" = 18, "una vez al día" = 30.

VI) Lugar_celular

"Cartera" = 0, "no contesta o bolsillo" = 1.

VII) Duerme_celular

"No" = 0, "no contesta o a veces" = 1, "sí" = 2.

VIII) Manos_libres_celular

"Sí" = 0, "no contesta o a veces" = 1, "no" = 2.

IX) Estudios_diag

"Sí" = 0, "no contesta o a veces" = 1.

X) AC[14] *preservativos*

"Sí" = 0, "no o no contesta" = 1.

XI) AC_oral:

"Sí" = 1, "no o no contesta" = 0.

XII) AC_inyectable

"Sí" = 1, "no o no contesta" = 0.

14 AC: anticonceptivo.

XIII) AC_DIU

"Sí" = 0, "no o no contesta" = 1.

XIV) AC_diafragma

"Sí" = 0, "no o no contesta" = 1.

XV) Cigarrillos_acum

Esta variable engloba numéricamente todos los valores recopilados. Se asume que aquel que fuma y no indica cantidad fuma 20 cigarrillos diarios. "Menos de 10" = 5, "entre 10 y 20" = 15, "más de 20" = 30, "más de 40" = 50. Y, además, que el fumador pasivo lo fue en la mitad de su edad.

Cigarrillos_acum = {fumador pasivo * [edad * (0,5) * cant horas exp/día * 2 cigarrillos * (0,5) * 365 días] + fumador activo * [cant años * cant cigarrillos / día * 365 días]}; donde "fumador pasivo" y "fumador activo" son variables binarias. A los fines del cálculo, al exfumador se lo trata como "fumador activo", ya que valen su antigüedad y frecuencia para generar una "carga acumulada".

XVI) Carnes_mes (en kg)

Para carnes rojas, consideramos 0,200 kg por porción para hombres y 0,180 kg para mujeres. Para carnes blancas, la porción se considera de 0,125 kg. Para pescados, la porción se considera en 0,170 kg y para mariscos 0,100 kg. Se tiene en cuenta la siguiente convención para el cálculo semanal de porciones: "no contesta" = 0 porciones, "menos de 1 vez por semana" = 0,5 porciones, "entre 1 y 2 veces por mes" = 1,5 / 4,35, "entre 1 y 2 veces por semana" = 1,5 porciones, "entre 2 y 3 veces por semana" = 2,5 porciones, "más de 3 veces por semana" = 4.

Para asignar la cantidad (en kg) al tipo de preparación se supone que cada preparación que tiene "sí" reparte el total uniformemente.

XVII) Aceite_mes (en cucharadas)

Se toma un criterio similar a lo anterior. Es decir: para asignar la cantidad (en cucharadas) al tipo de aceite se supone que cada tipo de aceite que tiene "sí" reparte el total del aceite uniformemente. La variable "otro_aceite" se elimina por presentar muy escasos valores y, de ellos, algunos numéricos y otros nominales.

XVIII) Embutidos y procesados

El valor de la porción es considerado en 0,050 kg.

XIX) Cereales, legumbres y hortalizas

El valor de la porción es considerado en 0,125 kg.

XX) Comida_baja_en_sal

Se hizo una variable booleana "comida_baja_en_sal" que tiene valor 1 solo cuando se indica como respuesta "bajas en sal" y tiene 0 "en otro caso", que incluye "no contesta", y "ns/nc".

XXI) Endulzantes

Se presentan las variables booleanas: usa_azúcar, usa_edulcorante, no_endulza. La variable "cucharadas de azúcar", que es numérica, se elimina porque sus valores son todos 0 (no varía).

XXII) Bebidas

Se anula "bebe agua embotellada" porque es redundante con la siguiente pregunta: "frecuencia de agua embotellada". Se considera que la porción de bebida blanca es de 50 cm3, la de té es de 150 cm3, la de aguas y bebidas no alcohólicas es de 200 cm3, la de vino es de 100 cm3. Los criterios de asignación de cantidades mensuales son similares a los anteriores. Para más de 3 l por semana, se consideraron 4 l. Además, para una botella de agua se consideraron 2 l. En el caso de "frecuencia de consumo de agua embotellada" = "no contesta" se asigna la respuesta "entre 3 y 5 vasos por día".

XXIII) Frec_tomografias

Se eliminó debido a que los encuestados no discriminaron en sus respuestas los estudios de este tipo realizados previos al diagnóstico de cáncer.

Codificación de variables de alimentos

1. **Cereales**: avena, arroz Yamani, trigo, mijo, quinua, hojuelas de maíz, *mix* con frutos secos.
2. **Legumbres**: soja, lentejas, garbanzos, otros granos.
3. **Comidas procesadas:** salchichas, embutidos, hamburguesas, fiambres.
4. **Bebidas sin alcohol:** aperitivos, gaseosas, aguas saborizadas y finamente gasificadas, soda.
5. **Bebidas saborizadas:** jugos en polvo o saborizantes con reconstitución (por ejemplo, granulados), aguas saborizadas.

No corresponde: la aparición de esta etiqueta en los resultados nos indica que nunca va a darse esa opción para la variable que se está analizando. Ej.: uso de anticonceptivos orales en hombres, control ginecológico en hombres.

Antecedentes patológicos

En el conjunto de datos, se diferencia si los antecedentes corresponden a madre, padre y demás familiares, y se codificaron de la siguiente manera:

Tabla I1. Antecedentes patológicos

AF_ Ld	Antecedente familiar línea directa	Hermano/a. Hijo/a
AF_Lm	Antecedente familiar línea materna	Madre. Abuelo/a. Tío/a
AF_Lp	Antecedente familiar línea paterna	Padre. Abuelo/a. Tío/a

Tabla I2. Codificación de los cánceres y enfermedades[15]

CÁNCERES CIE-10		Enfermedades	
NOMBRE DE LA VARIABLE COMÚN	**CÓDIGO**	**NOMBRE DE LA VARIABLE COMÚN**	**CÓDIGO**
Boca	C06	Alergias varias	Z71
Hipofaringe	C13	Alzheimer	F00
Cabeza y cuello	C14	Arritmia	I49
Esófago	C15	Artritis	E79
Estómago	C16	Artrosis	M15
Colon	C18	Asma	J45
Recto	C20	Cardiopatía congénita	I42
Hígado	C22	Celiaquía	D89
Vesícula	C23	Chagas Mazza	B57
Páncreas	C25	Diabetes mellitus tipo 2	E10
Pulmón	C34	Dislipidemia	E78
Mediastino	C38	Divertículos	K57
Laringe	C39	Epilepsia	G40
Hueso	C40	EPOC	J49
Osteosarcoma (C. vertebral)	C41	Fibromialgia	M79
Melanoma	C43	Gastritis crónica	K29
Piel	C44	Glaucoma	H40
Mesotelioma	C45	Hepatitis B	B16
Sarcoma de Kaposi	C46	Hiperinsulinemia	E89
Mama	C50	Hipertensión arterial	I10
Cuello de útero	C55	Hipertensión cardiaca	I11
Ovario	C56	Hipertensión pulmonar	J98
Cav. pélvica	C57	Hipertiroidismo	E05
Próstata	C61	Hipertrofia cardiaca	I51
Testículo	C62	Hipotiroidismo	E03
Epididimal (testículo)	C63	HIV	B20
Vejiga	C67	Lumbago	M50
Riñón	C68	Lupus	M32
Ojo y anexos	C69	Migraña	G43
Encéfalo	C71	Obesidad	E66
Medula ósea	C72	Osteoporosis	M80

15 Codificación de los cánceres y enfermedades, para el análisis por IA. Fuente: Elaboración propia, basado en CIE-10. Clasificación internacional de enfermedades. https://ais.paho.org/classifications/chapters/pdf/volume1.pdf.

Tiroides	C73	Parkinson	G20
Gl. suprarrenal	C74	Pólipos	D30
Linfoma no Hodking	C81	Psoriasis	L40
Mieloma	C90	Reflujo gastroesofágico	K21
Linfoma	C91	Síndrome de Hashimoto	E06
Leucemia	C95	Síndrome del colon irritable	F45
Tumor maligno de la vagina	C52	Vitíligo	L80

Listado N.° 1. Codificación de los medicamentos[16]

Grupo 01: analgésicos (alivio de dolores físicos).
Grupo 02: antiácidos y antiulcerosos (disminución de las secreciones gástricas).
Grupo 03: antihistamínicos (combatir efectos negativos de las reacciones alérgicas, loratadina, desloratadina cetirizina).
Grupo 04: antidiarreicos y laxantes (detener las diarreas, aliviar constipación).
Grupo 05: antiinfecciosos (dependiendo del agente infeccioso: antibióticos, antifúngicos, antivirales, antiparasitarios).
Grupo 06: antiinflamatorios (reducir las inflamaciones. Ej.: ibuprofeno, aspirina).
Grupo 07: antipiréticos (disminuir la temperatura corporal elevada, fiebre).
Grupo 08: antitusivos y mucolíticos (reducción de la tos no productiva).
Grupo 09: citostáticos (detienen o demoran el crecimiento de células cancerosas sin eliminarlas).

[16] Codificación de medicamentos. Elaboración propia, basado en PAHO. *Cuadro básico y catálogo de medicamentos*. 2018. México.

Grupo 10: psiquiátricos (inhibidor selectivo de la recapacitación de serotonina [ISRS], agentes antipsicóticos, estimulantes, antagonistas de alfa 2, agonistas de alfa 2, agonistas de la melatonina, antidepresivos, antiespasmódicos).

Grupo 11: antirreumatoideos (metrotexato, leflunomida, sulfasalazina, adalimubad, etanercep, infliximab, esteroides).

Grupo 12: anticonceptivos hormonales (orales, inyectables, dispositivo intrauterino, implantes, parches).

Grupo 13: antihipertensivos (inhibidor de la ECA, vasodilatadores, beta bloqueadores, inhibidores de los canales de calcio, nitroprusiato sódico, labetalol, fentolamina, urapidilo, furosemida, nitroglicerina sildenafil). Dentro de este grupo están: –***A:*** antihipertensivos neurológicos, –***B:*** antihipertensivos cardiológicos, –***C:*** antihipertensivos arteriales, –***D:*** antihipertensivos pulmonares.

Grupo 14: estatinas (reducen la absorción de triglicéridos: atorvastatina, rosuvastatina).

Grupo 15: antihiperglucemiantes (regulación de la glucosa en sangre: metformina, insulina inyectable pioglitazona).

Grupo 16: corticoides (reducción de la inflamación, efecto similar al de las hormonas suprarrenales: dexametasona).

Grupo 17: endocrinas (reemplazo hormonal: levotiroxina, paricalcitol, mesterolona, calcio).

Grupo 18: antiagregante plaquetario (previene trombosis arterial: aspirina, dipiridamol, clopidogrel, ticoplidina).

Grupo 19: antiarrítmicos (disminuir las taquiarritmias: amiodarona).

Grupo 20: suplementos vitamínicos y minerales (zinc, magnesio, calcio, otros).
Grupo 21: antidiuréticos (controlan el agua y la sal en el cuerpo: amlodip).

Anexo II. Variables del análisis 1 y 2

1. Listado de variables predictoras, primer análisis

Variables objetivo o target

variables_target= ['Dg_C81', 'Dg_C46', 'Dg_C56', 'Dg_C76', 'Dg_C25', 'Dg_C23', 'Dg_C50', 'Dg_C34', 'Dg_C20', 'Dg_C38', 'Dg_C41', 'Dg_C95', 'Dg_C16', 'Dg_C55', 'Dg_C06', 'Dg_C63', 'Dg_C61', 'Dg_C18', 'Dg_C43', 'Dg_C73', 'Dg_C62', 'Dg_C91', 'Dg_C57', 'Dg_C22', 'Dg_C44']

Se obtiene: variables = columnas [1:88]

variables = ['edad', 'hijos', 'peso', 'estatura', 'imc', 'Frec_RX_dental', 'Frec_RX', 'Fumador_Tiempo', 'Sexo_M', 'Ejercicios_hs_mes', 'Cama_solar', 'Pre_bronceado', 'Uso_sahumerios_mes', 'Celular_cartera', 'Celular_duerme', 'Celular_manos_libres', 'Estudios_diag', 'AC_Preservativos', 'AC_Inyectable', 'AC_diu', 'AC_diafragma', 'AC_oral', 'Cigarrillos_acum', 'Cant_comidas', 'carnesRojas_mes', 'carnesRojas_fritas_mes', 'carnesRojas_guiso_mes', 'carnesRojas_horno_mes', 'carnesRojas_plancha_mes', 'carnesRojas_parrilla_mes', 'carnesBlancas_mes', 'carnesBlancas_fritas_mes', 'carnesBlancas_guiso_mes', 'carnesBlancas_horno_mes', 'carnesBlancas_plancha_mes', 'carnesBlancas_parrilla_mes', 'Mariscos', 'mariscos_fritos_mes',

'mariscos_guiso_mes', 'mariscos_horno_mes', 'mariscos_plancha_mes', 'mariscos_parrilla_mes', 'Aceites', 'Ac_Oliva_', 'Ac_Mezcla_', 'Ac_Girasol_', 'Ac_Maiz_', 'Ac_noSabe_', 'Ac_en_Comidas', 'Pan', 'Facturas', 'Galletas', 'Golosinas', 'frutas_mes', 'bananas_mes', 'manzanas_mes', 'peras_mes', 'mandarinas_mes', 'naranjas_mes', 'otras_frutas_mes', 'Come_Fruta_pelada', 'Come_Fruta_con_cascara', 'Cereales_', 'Legumbres_', 'Hortalizas_', 'comida_baja_en_sal_', 'usa_azucar_', 'usa_edulcorante_', 'no_endulza_', 'te_mes_', 'Beb_blancas_mes_', 'vino_mes_', 'Beb_embotelladas_', 'Beb_azucaradas_', 'Beb_saborizadas_', 'Beb_light_', 'Beb_en_polvo_', 'agua_embotellada_mes', 'Ap_C18_', 'Ap_C34_', 'Ap_C44_', 'Ap_C50_', 'Ap_C55_', 'Ap_C61_', 'Ap_C91_', 'Ap_C95_', 'Ap_Cx_', 'Ap_CA_']

2. Listado de variables predictoras segundo análisis

Variables objetivo o target

variables_target= ['Dg_C81', 'Dg_C46', 'Dg_C90', 'Dg_C56', 'Dg_C76', 'Dg_C14', 'Dg_C25', 'Dg_C23', 'Dg_C50', 'Dg_C34', 'Dg_C39', 'Dg_C13', 'Dg_C38', 'Dg_C41', 'Dg_C95', 'Dg_C16', 'Dg_C55', 'Dg_C06', 'Dg_C63', 'Dg_C61', 'Dg_C18', 'Dg_C43', 'Dg_C15', 'Dg_C62', 'Dg_C91', 'Dg_C57', 'Dg_C22', 'Dg_C44']

Se obtiene: variables = columnas [1:849]

variables_indep= ['Sexo_Femenino', 'Sexo_Masculino', 'Edad', 'Estudios_Primario', 'Estudios_Secundario', 'Estudios_Terciario', 'Estudios_Universitario', 'E_Civil_casado', 'E_Civil_divorciado', 'E_Civil_soltero', 'E_Civil_viudo', 'Hijos', 'Peso', 'estatura', 'IMC', 'Talle', 'AP_B16', 'AP_B20', 'AP_B57', 'AP_C06', 'AP_C14', 'AP_C15', 'AP_C16', 'AP_C18', 'AP_C22',

'AP_C23', 'AP_C25', 'AP_C34', 'AP_C38', 'AP_C39', 'AP_C40', 'AP_C43', 'AP_C44', 'AP_C45', 'AP_C46', 'AP_C50', 'AP_C52', 'AP_C55', 'AP_C56', 'AP_C61', 'AP_C62', 'AP_C69', 'AP_C73', 'AP_C81', 'AP_C90', 'AP_C91', 'AP_C95', 'AP_D10', 'AP_D15', 'AP_D89', 'AP_E03', 'AP_E05', 'AP_E06', 'AP_E10', 'AP_E78', 'AP_E79', 'AP_E89', 'AP_F45', 'AP_G40', 'AP_G43', 'AP_H40', 'AP_I10', 'AP_I42', 'AP_I49', 'AP_I51', 'AP_J45', 'AP_J49', 'AP_J98', 'AP_K21', 'AP_K29', 'AP_K57', 'AP_L40', 'AP_L80', 'AP_M15', 'AP_M32', 'AP_M50', 'AP_M79', 'AP_M80', 'AP_Z71', 'Medicam_Grupo05', 'Medicam_Grupo16', 'Medicam_Grupo19', 'Medicam_Grupo03', 'Medicam_Grupo11', 'Medicam_Grupo01', 'Medicam_Grupo13D', 'Medicam_Grupo20', 'Medicam_Grupo21', 'Medicam_Grupo12', 'Medicam_Grupo15', 'Medicam_Grupo14', 'Medicam_Grupo10', 'Medicam_Grupo17', 'Medicam_Grupo02', 'Medicam_Grupo13B', 'Medicam_Grupo18', 'Medicam_Grupo06', 'Medicam_Grupo13C', 'Medicam_Grupo09', 'AF_C15', 'AF_Lm_C15', 'AF_Lp_C15', 'AF_Ld_C15', 'AF_Lm_M_C15', 'AF_Lm_F_C15', 'AF_Lp_M_C15', 'AF_Lp_F_C15', 'AF_Ld_M_C15', 'AF_Ld_F_C15', 'AF_C50', 'AF_Lm_C50', 'AF_Lp_C50', 'AF_Ld_C50', 'AF_Lm_M_C50', 'AF_Lm_F_C50', 'AF_Lp_M_C50', 'AF_Lp_F_C50', 'AF_Ld_M_C50', 'AF_Ld_F_C50', 'AF_K29', 'AF_Lm_K29', 'AF_Lp_K29', 'AF_Ld_K29', 'AF_Lm_M_K29', 'AF_Lm_F_K29', 'AF_Lp_M_K29', 'AF_Lp_F_K29', 'AF_Ld_M_K29', 'AF_Ld_F_K29', 'AF_C22', 'AF_Lm_C22', 'AF_Lp_C22', 'AF_Ld_C22', 'AF_Lm_M_C22', 'AF_Lm_F_C22', 'AF_Lp_M_C22', 'AF_Lp_F_C22', 'AF_Ld_M_C22', 'AF_Ld_F_C22', 'AF_C44', 'AF_Lm_C44', 'AF_Lp_C44', 'AF_Ld_C44', 'AF_Lm_M_C44', 'AF_Lm_F_C44', 'AF_Lp_M_C44', 'AF_Lp_F_C44', 'AF_Ld_M_C44', 'AF_Ld_F_C44', 'AF_C67', 'AF_Lm_C67', 'AF_Lp_C67', 'AF_Ld_C67', 'AF_Lm_M_C67', 'AF_Lm_F_C67', 'AF_Lp_M_C67', 'AF_Lp_F_C67', 'AF_Ld_M_C67',

'AF_Ld_F_C67', 'AF_C46', 'AF_Lm_C46', 'AF_Lp_C46', 'AF_Ld_C46', 'AF_Lm_M_C46', 'AF_Lm_F_C46', 'AF_Lp_M_C46', 'AF_Lp_F_C46', 'AF_Ld_M_C46', 'AF_Ld_F_C46', 'AF_E10', 'AF_Lm_E10', 'AF_Lp_E10', 'AF_Ld_E10', 'AF_Lm_M_E10', 'AF_Lm_F_E10', 'AF_Lp_M_E10', 'AF_Lp_F_E10', 'AF_Ld_M_E10', 'AF_Ld_F_E10', 'AF_C14', 'AF_Lm_C14', 'AF_Lp_C14', 'AF_Ld_C14', 'AF_Lm_M_C14', 'AF_Lm_F_C14', 'AF_Lp_M_C14', 'AF_Lp_F_C14', 'AF_Ld_M_C14', 'AF_Ld_F_C14', 'AF_C23', 'AF_Lm_C23', 'AF_Lp_C23', 'AF_Ld_C23', 'AF_Lm_M_C23', 'AF_Lm_F_C23', 'AF_Lp_M_C23', 'AF_Lp_F_C23', 'AF_Ld_M_C23', 'AF_Ld_F_C23', 'AF_C38', 'AF_Lm_C38', 'AF_Lp_C38', 'AF_Ld_C38', 'AF_Lm_M_C38', 'AF_Lm_F_C38', 'AF_Lp_M_C38', 'AF_Lp_F_C38', 'AF_Ld_M_C38', 'AF_Ld_F_C38', 'AF_C74', 'AF_Lm_C74', 'AF_Lp_C74', 'AF_Ld_C74', 'AF_Lm_M_C74', 'AF_Lm_F_C74', 'AF_Lp_M_C74', 'AF_Lp_F_C74', 'AF_Ld_M_C74', 'AF_Ld_F_C74', 'AF_C71', 'AF_Lm_C71', 'AF_Lp_C71', 'AF_Ld_C71', 'AF_Lm_M_C71', 'AF_Lm_F_C71', 'AF_Lp_M_C71', 'AF_Lp_F_C71', 'AF_Ld_M_C71', 'AF_Ld_F_C71', 'AF_C55', 'AF_Lm_C55', 'AF_Lp_C55', 'AF_Ld_C55', 'AF_Lm_M_C55', 'AF_Lm_F_C55', 'AF_Lp_M_C55', 'AF_Lp_F_C55', 'AF_Ld_M_C55', 'AF_Ld_F_C55', 'AF_K21', 'AF_Lm_K21', 'AF_Lp_K21', 'AF_Ld_K21', 'AF_Lm_M_K21', 'AF_Lm_F_K21', 'AF_Lp_M_K21', 'AF_Lp_F_K21', 'AF_Ld_M_K21', 'AF_Ld_F_K21', 'AF_C61', 'AF_Lm_C61', 'AF_Lp_C61', 'AF_Ld_C61', 'AF_Lm_M_C61', 'AF_Lm_F_C61', 'AF_Lp_M_C61', 'AF_Lp_F_C61', 'AF_Ld_M_C61', 'AF_Ld_F_C61', 'AF_I10', 'AF_Lm_I10', 'AF_Lp_I10', 'AF_Ld_I10', 'AF_Lm_M_I10', 'AF_Lm_F_I10', 'AF_Lp_M_I10', 'AF_Lp_F_I10', 'AF_Ld_M_I10', 'AF_Ld_F_I10', 'AF_C18', 'AF_Lm_C18', 'AF_Lp_C18', 'AF_Ld_C18', 'AF_Lm_M_C18',

'AF_Lm_F_C18', 'AF_Lp_M_C18', 'AF_Lp_F_C18', 'AF_Ld_M_C18', 'AF_Ld_F_C18', 'AF_C34', 'AF_Lm_C34', 'AF_Lp_C34', 'AF_Ld_C34', 'AF_Lm_M_C34', 'AF_Lm_F_C34', 'AF_Lp_M_C34', 'AF_Lp_F_C34', 'AF_Ld_M_C34', 'AF_Ld_F_C34', 'AF_C68', 'AF_Lm_C68', 'AF_Lp_C68', 'AF_Ld_C68', 'AF_Lm_M_C68', 'AF_Lm_F_C68', 'AF_Lp_M_C68', 'AF_Lp_F_C68', 'AF_Ld_M_C68', 'AF_Ld_F_C68', 'AF_C16', 'AF_Lm_C16', 'AF_Lp_C16', 'AF_Ld_C16', 'AF_Lm_M_C16', 'AF_Lm_F_C16', 'AF_Lp_M_C16', 'AF_Lp_F_C16', 'AF_Ld_M_C16', 'AF_Ld_F_C16', 'AF_M32', 'AF_Lm_M32', 'AF_Lp_M32', 'AF_Ld_M32', 'AF_Lm_M_M32', 'AF_Lm_F_M32', 'AF_Lp_M_M32', 'AF_Lp_F_M32', 'AF_Ld_M_M32', 'AF_Ld_F_M32', 'AF_D89', 'AF_Lm_D89', 'AF_Lp_D89', 'AF_Ld_D89', 'AF_Lm_M_D89', 'AF_Lm_F_D89', 'AF_Lp_M_D89', 'AF_Lp_F_D89', 'AF_Ld_M_D89', 'AF_Ld_F_D89', 'AF_ C25', 'AF_Lm_ C25', 'AF_Lp_ C25', 'AF_Ld_ C25', 'AF_Lm_M_ C25', 'AF_Lm_F_ C25', 'AF_Lp_M_ C25', 'AF_Lp_F_ C25', 'AF_Ld_M_ C25', 'AF_Ld_F_ C25', 'AF_J45', 'AF_Lm_J45', 'AF_Lp_J45', 'AF_Ld_J45', 'AF_Lm_M_J45', 'AF_Lm_F_J45', 'AF_Lp_M_J45', 'AF_Lp_F_J45', 'AF_Ld_M_J45', 'AF_Ld_F_J45', 'AF_G20', 'AF_Lm_G20', 'AF_Lp_G20', 'AF_Ld_G20', 'AF_Lm_M_G20', 'AF_Lm_F_G20', 'AF_Lp_M_G20', 'AF_Lp_F_G20', 'AF_Ld_M_G20', 'AF_Ld_F_G20', 'AF_C62', 'AF_Lm_C62', 'AF_Lp_C62', 'AF_Ld_C62', 'AF_Lm_M_C62', 'AF_Lm_F_C62', 'AF_Lp_M_C62', 'AF_Lp_F_C62', 'AF_Ld_M_C62', 'AF_Ld_F_C62', 'AF_C95', 'AF_Lm_C95', 'AF_Lp_C95', 'AF_Ld_C95', 'AF_Lm_M_C95', 'AF_Lm_F_C95', 'AF_Lp_M_C95', 'AF_Lp_F_C95', 'AF_Ld_M_C95', 'AF_Ld_F_C95', 'AF_C43', 'AF_Lm_C43', 'AF_Lp_C43', 'AF_Ld_C43', 'AF_Lm_M_C43', 'AF_Lm_F_C43', 'AF_Lp_M_C43', 'AF_Lp_F_C43', 'AF_Ld_M_C43', 'AF_Ld_F_C43', 'AF_I11', 'AF_Lm_I11', 'AF_Lp_I11',

'AF_Ld_I11', 'AF_Lm_M_I11', 'AF_Lm_F_I11', 'AF_Lp_M_I11', 'AF_Lp_F_I11', 'AF_Ld_M_I11', 'AF_Ld_F_I11', 'AF_C72', 'AF_Lm_C72', 'AF_Lp_C72', 'AF_Ld_C72', 'AF_Lm_M_C72', 'AF_Lm_F_C72', 'AF_Lp_M_C72', 'AF_Lp_F_C72', 'AF_Ld_M_C72', 'AF_Ld_F_C72', 'AF_F00', 'AF_Lm_F00', 'AF_Lp_F00', 'AF_Ld_F00', 'AF_Lm_M_F00', 'AF_Lm_F_F00', 'AF_Lp_M_F00', 'AF_Lp_F_F00', 'AF_Ld_M_F00', 'AF_Ld_F_F00', 'AF_C40', 'AF_Lm_C40', 'AF_Lp_C40', 'AF_Ld_C40', 'AF_Lm_M_C40', 'AF_Lm_F_C40', 'AF_Lp_M_C40', 'AF_Lp_F_C40', 'AF_Ld_M_C40', 'AF_Ld_F_C40', 'AF_C90', 'AF_Lm_C90', 'AF_Lp_C90', 'AF_Ld_C90', 'AF_Lm_M_C90', 'AF_Lm_F_C90', 'AF_Lp_M_C90', 'AF_Lp_F_C90', 'AF_Ld_M_C90', 'AF_Ld_F_C90', 'AF_C06', 'AF_Lm_C06', 'AF_Lp_C06', 'AF_Ld_C06', 'AF_Lm_M_C06', 'AF_Lm_F_C06', 'AF_Lp_M_C06', 'AF_Lp_F_C06', 'AF_Ld_M_C06', 'AF_Ld_F_C06', 'AF_C39', 'AF_Lm_C39', 'AF_Lp_C39', 'AF_Ld_C39', 'AF_Lm_M_C39', 'AF_Lm_F_C39', 'AF_Lp_M_C39', 'AF_Lp_F_C39', 'AF_Ld_M_C39', 'AF_Ld_F_C39', 'AF_C73', 'AF_Lm_C73', 'AF_Lp_C73', 'AF_Ld_C73', 'AF_Lm_M_C73', 'AF_Lm_F_C73', 'AF_Lp_M_C73', 'AF_Lp_F_C73', 'AF_Ld_M_C73', 'AF_Ld_F_C73', 'AF_E66', 'AF_Lm_E66', 'AF_Lp_E66', 'AF_Ld_E66', 'AF_Lm_M_E66', 'AF_Lm_F_E66', 'AF_Lp_M_E66', 'AF_Lp_F_E66', 'AF_Ld_M_E66', 'AF_Ld_F_E66', 'AF_E78', 'AF_Lm_E78', 'AF_Lp_E78', 'AF_Ld_E78', 'AF_Lm_M_E78', 'AF_Lm_F_E78', 'AF_Lp_M_E78', 'AF_Lp_F_E78', 'AF_Ld_M_E78', 'AF_Ld_F_E78', 'AF_C25', 'AF_Lm_C25', 'AF_Lp_C25', 'AF_Ld_C25', 'AF_Lm_M_C25', 'AF_Lm_F_C25', 'AF_Lp_M_C25', 'AF_Lp_F_C25', 'AF_Ld_M_C25', 'AF_Ld_F_C25', 'AF_C91', 'AF_Lm_C91', 'AF_Lp_C91', 'AF_Ld_C91', 'AF_Lm_M_C91', 'AF_Lm_F_C91', 'AF_Lp_M_C91', 'AF_Lp_F_C91', 'AF_Ld_M_C91',

'AF_Ld_F_C91', 'ejercicios_hs_mes', 'ejercicios_hace', 'ejerc_frec_Entre 1 y 2 veces por semana', 'ejerc_frec_Entre dos y 4 veces por mes', 'ejerc_frec_NC', 'ejerc_frec_Nunca', 'ejerc_frec_Todos los dias', 'ejerc_frec_Una vez por mes', 'ejerc_frec_Una vez por semana', 'cama_solar_frecuencia', 'cama_solar_usa', 'cama_solar_+4_veces', 'cama_solar_0_veces', 'cama_solar_1_vez', 'cama_solar_NC', 'cama_solar_e_2_y_4_veces', 'sahumerios_frec_semana', 'sahumerios_usa', 'sahumerios_semana__+3_veces', 'sahumerios_semana__0_veces', 'sahumerios_semana__7_veces', 'sahumerios_semana__NC', 'sahumerios_semana__e_1_y_3_veces', 'prebroncea', 'prebroncea_A_veces', 'prebroncea_Casi_nunca', 'prebroncea_Casi_siempre', 'prebroncea_No contesta', 'prebroncea_Nunca', 'prebroncea_Siempre', 'toma_sol', 'toma_sol_NC', 'toma_sol_Nunca', 'toma_sol_antes_de_11hs', 'toma_sol_entre_12hs_y_15hs', 'toma_sol_luego_de_16hs', 'bloqueador_usa', 'bloqueador_A_veces', 'bloqueador_Casi_nunca', 'bloqueador_NC', 'bloqueador_NaN', 'bloqueador_No', 'bloqueador_Siempre', 'celular_lleva_Bolsillo', 'celular_lleva_Cartera', 'celular_lleva_NC', 'celular_mesa_luz', 'celular_mesa_luz_A_veces', 'celular_mesa_luz_NC', 'celular_mesa_luz_No', 'celular_mesa_luz_Si', 'celular_manos_libres', 'celular_manos_libres_A_Veces', 'celular_manos_libres_NC', 'celular_manos_libres_No', 'celular_manos_libres_Si', 'comidas_cant', 'carne_roja_mes', 'carne_roja_come', 'carne_roja_+3_veces_semana', 'carne_roja_NC', 'carne_roja_Nunca', 'carne_roja_a_diario', 'carne_roja_e_1_y_2_veces_mes', 'carne_roja_e_1_y_2_veces_semana', 'carne_roja_e_2_y_3_veces_semana', 'carne_roja_hasta_1_vez_semana', 'carne_blanca_mes', 'carne_blanca_come', 'carne_blanca_+3_veces_semana', 'carne_blanca_NC', 'carne_blanca_Nunca', 'carne_blanca_a_diario',

'carne_blanca_e_1_y_2_veces_mes', 'carne_blanca_e_1_y_2_veces_semana', 'carne_blanca_e_2_y_3_veces_semana', 'pescados_y_mariscos_mes', 'pescados_y_mariscos_come', 'pescados_y_mariscos_+3_veces_semana', 'pescados_y_mariscos_NC', 'pescados_y_mariscos_Nunca', 'pescados_y_mariscos_a_diario', 'pescados_y_mariscos_e_1_y_2_veces_mes', 'pescados_y_mariscos_e_1_y_2_veces_semana', 'pescados_y_mariscos_e_2_y_3_veces_semana', 'comidas_procesadas_mes', 'comidas_procesadas_come', 'comidas_procesadas_+3_veces_semana', 'comidas_procesadas_NC', 'comidas_procesadas_Nunca', 'comidas_procesadas_a_diario', 'comidas_procesadas_e_1_y_2_veces_mes', 'comidas_procesadas_e_1_y_2_veces_semana', 'comidas_procesadas_e_2_y_3_veces_semana', 'come_encurtidos_mes', 'encurtidos_come', 'encurtidos_come_+3_veces_semana', 'encurtidos_come_NC', 'encurtidos_come_Nunca', 'encurtidos_come_a_diario', 'encurtidos_come_e_1_y_2_veces_mes', 'encurtidos_come_e_1_y_2_veces_semana', 'encurtidos_come_e_2_y_3_veces_semana', 'come_frutas_mes', 'frutas_come', 'frutas_come_+3_veces_semana', 'frutas_come_NC', 'frutas_come_Nunca', 'frutas_come_a_diario', 'frutas_come_e_1_y_2_veces_mes', 'frutas_come_e_1_y_2_veces_semana', 'frutas_come_e_2_y_3_veces_semana', 'banana_come_No', 'banana_come_Si', 'manzana_come_No', 'manzana_come_Si', 'pera_come_No', 'pera_come_Si', 'mandarina_come_No', 'mandarina_come_Si', 'naranja_come_No', 'naranja_come_Si', 'fruta_come_Con cáscara', 'fruta_come_No contesta', 'fruta_come_Pelada', 'otras_frutas_come_NO', 'otras_frutas_come_SI', 'cereales_come_mes', 'cereales_come', 'cereales_come_+3_veces_semana', 'cereales_come_NC', 'cerea-

les_come_Nunca', 'cereales_come_a_diario', 'cereales_come_e_1_y_2_veces_mes', 'cereales_come_e_1_y_2_veces_semana', 'cereales_come_e_2_y_3_veces_semana', 'legumbres_come_mes', 'legumbres_come', 'legumbres_come_+3_veces_semana', 'legumbres_come_NC', 'legumbres_come_Nunca', 'legumbres_come_a_diario', 'legumbres_come_e_1_y_2_veces_mes', 'legumbres_come_e_1_y_2_veces_semana', 'legumbres_come_e_2_y_3_veces_semana', 'hortalizas_come_mes', 'hortalizas_come', 'hortalizas_come_+3_veces_semana', 'hortalizas_come_NC', 'hortalizas_come_Nunca', 'hortalizas_come_a_diario', 'hortalizas_come_e_1_y_2_veces_mes', 'hortalizas_come_e_1_y_2_veces_semana', 'hortalizas_come_e_2_y_3_veces_semana', 'toma_Te_tazas_mes', 'Te_toma', 'Te_toma_+3_veces_semana', 'Te_toma_NC', 'Te_toma_Nunca', 'Te_toma_a_diario', 'Te_toma_e_1_y_2_veces_mes', 'Te_toma_e_1_y_2_veces_semana', 'Te_toma_e_2_y_3_veces_semana', 'bebidas_blancas_Lts_mes', 'bebidas_blancas_toma', 'bebidas_blancas_toma_+3_vasos_semana', 'bebidas_blancas_toma_1_vaso_semana', 'bebidas_blancas_toma_NC', 'bebidas_blancas_toma_Nunca', 'bebidas_blancas_toma_e_1_y_2_vasos_dia', 'bebidas_blancas_toma_e_1_y_3_vasos_mes', 'bebidas_blancas_toma_e_2_y_3_vasos_semana', 'bebidas_blancas_toma_e_3_y_5_vasos_dia', 'vino_toma_vasos_mes', 'vino_toma', 'vino_toma_+3_vasos_semana', 'vino_toma_NC', 'vino_toma_Nunca', 'vino_toma_e_1_y_2_vasos_dia', 'vino_toma_e_1_y_2_vasos_mes', 'vino_toma_e_2_y_3_vasos_semana',

'vino_toma_e_3_y_5_vasos_dia', 'bebidas_sin_alcohol_toma_Lts_mes', 'bebidas_sin_alcohol_toma', 'bebidas_sin_alcohol_toma_+3_lts_semana', 'bebidas_sin_alcohol_toma_1_botella_semana', 'bebidas_sin_alcohol_toma_NC', 'bebidas_sin_alcohol_toma_e_1_y_2_vasos_dia', 'bebidas_sin_alcohol_toma_e_1_y_2_vasos_semana', 'bebidas_sin_alcohol_toma_e_2_y_3_vasos_semana', 'bebidas_sin_alcohol_toma_e_3_y_5_vasos_dia', 'carnes_rojas_fritas_come', 'carnes_rojas_fritas_come_No', 'carnes_rojas_fritas_come_Si', 'carnes_rojas_guisadas_come', 'carnes_rojas_guisadas_come_No', 'carnes_rojas_guisadas_come_Si', 'carnes_rojas_horneadas_come', 'carnes_rojas_horneadas_come_No', 'carnes_rojas_horneadas_come_Si', 'carnes_rojas_a_la_plancha_come', 'carnes_rojas_a_la_plancha_come_No', 'carnes_rojas_a_la_plancha_come_Si', 'carnes_rojas_a_la_parrilla_come', 'carnes_rojas_a_la_parrilla_come_No', 'carnes_rojas_a_la_parrilla_come_Si', 'carnes_blancas_fritas_come', 'carnes_blancas_fritas_come_No', 'carnes_blancas_fritas_come_Si', 'carnes_blancas_guisadas_come', 'carnes_blancas_guisadas_come_No', 'carnes_blancas_guisadas_come_Si', 'carnes_blancas_horneadas_come', 'carnes_blancas_horneadas_come_No', 'carnes_blancas_horneadas_come_Si', 'carnes_blancas_a_la_plancha_come', 'carnes_blancas_a_la_plancha_come_No', 'carnes_blancas_a_la_plancha_come_Si', 'carnes_blancas_a_la_parrilla_come', 'carnes_blancas_a_la_parrilla_come_No', 'carnes_blancas_a_la_parrilla_come_Si', 'pescados_y_mariscos_fritos_come', 'pescados_y_mariscos_fritos_come_No', 'pescados_y_mariscos_fritos_come_Si', 'pescados_y_mariscos_guisados_come', 'pescados_y_mariscos_guisa-

dos_come_No', 'pescados_y_mariscos_guisados_come_Si', 'pescados_y_mariscos_horneados_come', 'pescados_y_mariscos_horneados_come_No', 'pescados_y_mariscos_horneados_come_Si', 'pescados_y_mariscos_a_la_plancha_come', 'pescados_y_mariscos_a_la_plancha_come_No', 'pescados_y_mariscos_a_la_plancha_come_Si', 'pescados_y_mariscos_a_la_parrilla_come', 'pescados_y_mariscos_a_la_parrilla_come_No', 'pescados_y_mariscos_a_la_parrilla_come_Si', 'manteca_come', 'manteca_come_No', 'manteca_come_Si', 'margarina_come', 'margarina_come_No', 'margarina_come_Si', 'grasa_come', 'grasa_come_No', 'grasa_come_Si', 'aceite_Oliva_come', 'aceite_Oliva_come_No', 'aceite_Oliva_come_Si', 'aceite_mezcla_come', 'aceite_mezcla_come_No', 'aceite_mezcla_come_Si', 'aceite_Girasol_come', 'aceite_Girasol_come_No', 'aceite_Girasol_come_Si', 'aceite_Maiz_come', 'aceite_Maiz_come_No', 'aceite_Maiz_come_Si', 'aceite_no_sabe_tipo_come', 'aceite_no_sabe_tipo_come_No', 'aceite_no_sabe_tipo_come_Si', 'aceite_otro_tipo_come', 'aceite_crudo', 'aceite_crudo_come', 'aceite_cocido', 'aceite_cocido_come', 'pan_cant', 'pan_come', 'facturas_cant', 'facturas_come', 'galletas_cant', 'galletas_come', 'golosinas_cant', 'golosinas_come', 'endulza_con_azucar_No', 'endulza_con_azucar_Si', 'endulza_con_edulcorante_No', 'endulza_con_edulcorante_Si', 'endulza_no_No', 'endulza_no_Si', 'Sal_en_comidas_Ns/Nc', 'Sal_en_comidas_mucha', 'Sal_en_comidas_poca', 'bebidas_azucaradas_toma_No', 'bebidas_azucaradas_toma_Si', 'bebidas_saborizadas_toma_No', 'bebidas_saborizadas_toma_Si', 'bebidas_light_toma_No', 'bebidas_light_toma_Si', 'bebidas_en_polvo_toma_No', 'bebidas_en_polvo_toma_Si', 'RX_cant', 'RX_seHizo', 'RX_Den-

tal_cant', 'RX_Dental_seHizo', 'control_gineco_mas_de_24_meses', 'control_gineco_menos_de_06_meses', 'control_gineco_menos_de_12_meses', 'control_gineco_menos_de_24_meses', 'control_gineco_no_corresponde', 'control_gineco_ns_nc', 'control_gineco_nunca', 'papa_nicolaou_mas_de_24_meses', 'papa_nicolaou_menos_de_06_meses', 'papa_nicolaou_menos_de_12_meses', 'papa_nicolaou_menos_de_24_meses', 'papa_nicolaou_no_corresponde', 'papa_nicolaou_ns_nc', 'papa_nicolaou_nunca', 'ac_preservativos_no', 'ac_preservativos_si', 'ac_oral_no', 'ac_oral_no_corresponde', 'ac_oral_si', 'ac_inyectable_no', 'ac_inyectable_no_corresponde', 'ac_inyectable_si', 'ac_diu_no', 'ac_diu_no_corresponde', 'ac_diu_si', 'ac_diafragma_no', 'ac_diafragma_no_corresponde', 'ac_diafragma_si', 'fumador_no', 'fumador_si', 'ex_fumador_no', 'ex_fumador_si', 'fuma_cigarrillos_acumulados', 'fuma_cant_dia_entre_0_y_10', 'fuma_cant_dia_entre_10_y_20', 'fuma_cant_dia_entre_20_y_40', 'fuma_cant_dia_error', 'fuma_cant_dia_mas_de_40', 'fuma_cant_dia_ninguno', 'fumador_Pasivo_no', 'fumador_Pasivo_si', 'fumador_pasivo_tiempo_0h', 'fumador_pasivo_tiempo_e_0h_y_1h', 'fumador_pasivo_tiempo_e_1h_y_2h', 'fumador_pasivo_tiempo_mas_de_2h', 'fumador_pasivo_tiempo_ns_nc']

Anexo III. Aprendizaje no supervisado

1. Reducción de dimensionalidad

K-means con variables seleccionadas por Entropía

Las tablas de linfoma y leucemia muestran que el cruce de las variables seleccionadas por el algoritmo siempre da un valor bajo.

Tabla AIII 1

Diagnóstico C-91 (linfoma)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
carnesRojas_ parrilla_mes; Aceite	carnesRojas_ parrilla_mes	Aceite	Bajo	Bajo
Sol_hs_mes; Aceite	Sol_hs_mes	Aceite	Bajo	Bajo

Tabla AIII 2

Diagnóstico C-95 (leucemia)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
('carnesRojas_ mes', 'Sol_hs_mes')	carnesRojas_mes	Sol_hs_mes	Bajo	Bajo
('carnesRojas_ mes', 'Cigarrillos_acum')	carnesRojas_mes	Cigarrillos_acum	Bajo	Bajo
('carnesRojas_ mes','carnesRojas_ parrilla_mes')	carnesRojas_mes	carnesRojas_ parrilla_mes	Bajo	Bajo
('Sol_hs_mes', 'Cigarrillos_acum')	Sol_hs_mes	cigarrillos_acum	Bajo	Bajo

Mientras que no hubo resultados para los siguientes diagnósticos: *C-18 (colon)*, *C-34 (pulmón)*, *C-44 (piel)*, *C-50 (mama) y C-55 (cuello de útero)*, *C-61 (próstata).*

K-means con variables seleccionadas por correlación positiva

A continuación, se presenta un listado por cada tipo de cáncer. En este listado, se presentan combinaciones de variables seleccionadas por correlación positiva, donde todos los casos diagnosticados como positivos están totalmente agrupados dentro de un mismo tipo de cáncer.

Diagnóstico C-18 (colon), diagnóstico C-34 (pulmón), diagnóstico C-44 (piel), diagnóstico C-50 (mama), diagnóstico C-55 (útero), diagnóstico C-61 (próstata), diagnóstico C-91 (linfoma), diagnóstico C-95 (leucemia): No hay coincidencias.

K-means con variables seleccionadas por correlación negativa

A continuación, se presenta un listado por cada tipo de cáncer. En este listado, se presentan combinaciones de variables seleccionadas por correlación negativa, donde todos los casos diagnosticados como positivos están totalmente incluidos dentro de un mismo tipo de cáncer.

Cuando la gráfica muestra los conjuntos separados y sin ninguna cercanía, estamos frente a variables con respuestas dicotómicas.

Tabla AIII 3

Diagnóstico C-18 (colon)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
AC_oral', 'otrasFrutas_peladas_mes'	'AC_oral'	otrasFrutas_peladas_mes'	Bajo	Bajo
'AC_oral', 'AC_diu'	'AC_oral'	'AC_diu'	Bajo	Bajo
'AC_diu', 'otrasFrutas_peladas_mes'	AC_diu'	'otrasFrutas_peladas_mes'	Bajo	Bajo

Tabla AIII 4

Diagnóstico C-34 (pulmón)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
AC_oral', 'carnesBlancas_fritas_mes'	'AC_oral'	carnesBlancas_fritas_mes'	Bajo	Bajo
'AC_oral', 'Beb_blancas_mes'	'AC_oral'	'Beb_blancas_mes'	Bajo	Bajo
AC_oral', 'mariscos_guiso_mes'	'AC_oral'	'mariscos_guiso_mes'	Bajo	Bajo
Ejercicios_hs_mes', 'carnesBlancas_fritas_mes'	'Ejercicios_hs_mes'	carnesBlancas_fritas_mes'	Bajo	Bajo
Ejercicios_hs_mes', 'Beb_blancas_mes'	'Ejercicios_hs_mes'	'Beb_blancas_mes'	Bajo	Bajo
Ejercicios_hs_mes', 'mariscos_guiso_mes'	'Ejercicios_hs_mes'	'mariscos_guiso_mes'	Bajo	Bajo
carnesBlancas_fritas_mes', 'Beb_blancas_mes'	'carnesBlancas_fritas_mes'	'Beb_blancas_mes'	Bajo	Bajo
carnesBlancas_fritas_mes', 'mariscos_guiso_mes'	'carnesBlancas_fritas_mes'	'mariscos_guiso_mes'	Bajo	Bajo
Beb_blancas_mes', 'mariscos_guiso_mes'	'Beb_blancas_mes'	'mariscos_guiso_mes'	Bajo	Bajo
'Beb_blancas_mes', 'mariscos_guiso_mes'	'Beb_blancas_mes'	'mariscos_guiso_mes'	Bajo	Bajo

Tabla AIII 5

Diagnóstico C-44 (piel)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
Sol_hs_mes', 'Legumbres'	'Sol_hs_mes'	'Legumbres'	Bajo	Bajo
Sol_hs_mes', 'Golosinas'	'Sol_hs_mes'	'Golosinas'	Bajo	Bajo
'Sol_hs_mes', 'manzanas_ peladas_mes'	'Sol_hs_mes'	'manzanas_ peladas_mes	Bajo	Bajo
carnesBlancas_ horno_mes', 'Golosinas'	'carnesBlancas_ horno_mes'	'Golosinas'	Bajo	Bajo
carnesBlancas_ horno_mes', 'manzanas_ peladas_mes'	'carnesBlancas_ horno_mes'	'manzanas_ peladas_mes	Bajo	Bajo
Legumbres', 'Facturas'	'Legumbres'	'Facturas'	Bajo	Bajo
Legumbres', 'frutas_peladas_ mes'	'Legumbres'	'frutas_peladas_ mes'	Bajo	Bajo
'Legumbres', 'Ejercicios_hs_ mes'	'Legumbres'	Ejercicios_hs_mes'	Bajo	Bajo
'Legumbres', 'Golosinas'	'Legumbres'	'Golosinas'	Bajo	Bajo
Facturas', 'Ejercicios_hs_ mes'	'Facturas'	Ejercicios_hs_mes'	Bajo	Bajo
'Facturas', 'Golosinas'	'Facturas'	'Golosinas'	Bajo	Bajo
'frutas_peladas_ mes', 'Golosinas'	'frutas_peladas_ mes'	'Golosinas'	Bajo	Bajo
'Ejercicios_hs_ mes', 'Golosinas'	Ejercicios_hs_mes'	'Golosinas'	Bajo	Bajo
Ejercicios_hs_mes', 'Frec_RX_dental'	Ejercicios_hs_mes'	'Frec_RX_dental'	Bajo	Bajo
'Golosinas', 'manzanas_ peladas_mes'	'Golosinas'	'manzanas_peladas_ mes	Bajo	Bajo

Las tablas de los cánceres de colon, pulmón y piel muestran que el cruce de las variables seleccionadas por el algoritmo siempre da un valor bajo.

No hubo resultados para el diagnóstico C-55 (cuello de útero).

Tabla AIII 6

Diagnóstico C-91 linfoma				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
Aceites', 'Sol_hs_mes'	'Aceites'	'Sol_hs_mes'	Bajo	Bajo
Aceites', 'Estudios_diag'	'Aceites'	'Estudios_diag'	Bajo	Bajo
Aceites', 'Ac_en_Comidas'	'Aceites'	'Ac_en_Comidas'	Bajo	Bajo
Aceites', 'Legumbres'	'Aceites'	'Legumbres'	Bajo	Bajo
Aceites', 'frutas_peladas_mes'	'Aceites'	'frutas_peladas_mes'	Bajo	Bajo
'Aceites', 'Frec_RX'	'Aceites'	'Frec_RX'	Bajo	Bajo
Aceites', 'AC_oral'	'Aceites'	'AC_oral'	Bajo	Bajo
Aceites', 'manzanas_peladas_mes'	'Aceites'	'manzanas_peladas_mes'	Bajo	Bajo
Aceites', 'carnesRojas_parrilla_mes'	'Aceites'	'carnesRojas_parrilla_mes'	Bajo	Bajo
Aceites', 'Ac_Oliva_'	'Aceites'	'Ac_Oliva_'	Bajo	Bajo
Aceites', 'peras_peladas_mes'	'Aceites'	'peras_peladas_mes'	Bajo	Bajo
Aceites', 'Ac_Girasol_'	'Aceites'	'Ac_Girasol_'	Bajo	Bajo
Estudios_diag', 'Sol_hs_mes'	'Estudios_diag'	'Sol_hs_mes'	Bajo	Bajo
Estudios_diag', 'Legumbres'	'Estudios_diag'	'Legumbres'	Bajo	Bajo
Estudios_diag', 'frutas_peladas_mes'	'Estudios_diag'	'frutas_peladas_mes'	Bajo	Bajo
Estudios_diag', 'Ac_en_Comidas'	'Estudios_diag'	'Ac_en_Comidas'	Bajo	Bajo
Estudios_diag', 'AC_oral'	'Estudios_diag'	'AC_oral'	Bajo	Bajo

Estudios_diag', 'manzanas_peladas_mes'	'Estudios_diag'	'manzanas_peladas_mes'	Bajo	Bajo
Estudios_diag', 'Frec_RX'	'Estudios_diag'	'Frec_RX'	Bajo	Bajo
'Estudios_diag', 'Ac_Oliva_'	'Estudios_diag'	'Ac_Oliva_'	Bajo	Bajo
'Estudios_diag', 'peras_peladas_mes'	'Estudios_diag'	'peras_peladas_mes'	Bajo	Bajo
'Estudios_diag', 'Ac_Girasol_'	'Estudios_diag'	'Ac_Girasol_'	Bajo	Bajo
'Sol_hs_mes', 'Legumbres'	'Sol_hs_mes',	'Legumbres'	Bajo	Bajo
'Sol_hs_mes', 'manzanas_peladas_mes'	'Sol_hs_mes',	'manzanas_peladas_mes'	Bajo	Bajo
'Sol_hs_mes', 'Frec_RX'	'Sol_hs_mes',	'Frec_RX'	Bajo	Bajo
'Sol_hs_mes', 'AC_oral'	'Sol_hs_mes',	'AC_oral'	Bajo	Bajo
'Ac_en_Comidas', 'Legumbres'	'Ac_en_Comidas'	'Legumbres'	Bajo	Bajo
'Ac_en_Comidas', 'AC_oral'	'Ac_en_Comidas'	'AC_oral'	Bajo	Bajo
'Ac_en_Comidas', 'manzanas_peladas_mes'	'Ac_en_Comidas'	'manzanas_peladas_mes'	Bajo	Bajo
'Ac_en_Comidas', 'Frec_RX'	'Ac_en_Comidas'	'Frec_RX'	Bajo	Bajo
'Ac_en_Comidas', 'Ac_Oliva_'	'Ac_en_Comidas'	'Ac_Oliva_'	Bajo	Bajo
'Ac_en_Comidas', 'Ac_Girasol_'	'Ac_en_Comidas'	'Ac_Girasol_'	Bajo	Bajo
'frutas_peladas_mes', 'AC_oral'	'frutas_peladas_mes'	'AC_oral'	Bajo	Bajo
'frutas_peladas_mes', 'Legumbres'	'frutas_peladas_mes'	'Legumbres'	Bajo	Bajo
'frutas_peladas_mes','Frec_RX'	'frutas_peladas_mes'	'Frec_RX'	Bajo	Bajo
'frutas_peladas_mes', 'manzanas_peladas_mes'	'frutas_peladas_mes'	'manzanas_peladas_mes'	Bajo	Bajo
'Legumbres', 'AC_oral'	'frutas_peladas_mes'	'AC_oral'	Bajo	Bajo
'frutas_peladas_mes','peras_peladas_mes'	'frutas_peladas_mes'	'peras_peladas_mes'	Bajo	Bajo

'Legumbres', 'Frec_RX'	'Legumbres'	'Frec_RX'	Bajo	Bajo
'Legumbres', 'Ac_Oliva_'	'Legumbres'	'Ac_Oliva_'	Bajo	Bajo
'Legumbres', 'manzanas_peladas_mes'	'Legumbres'	'manzanas_peladas_mes'	Bajo	Bajo
'Legumbres', 'Ac_Girasol_'	'Legumbres'	'Ac_Girasol_'	Bajo	Bajo
'Legumbres', 'peras_peladas_mes'	'Legumbres'	'peras_peladas_mes'	Bajo	Bajo
'AC_oral', 'Frec_RX'	'AC_oral'	'Frec_RX'	Bajo	Bajo
'AC_oral', 'manzanas_peladas_mes'	'AC_oral'	'manzanas_peladas_mes'	Bajo	Bajo
'AC_oral', 'Ac_Oliva_'	'AC_oral'	'Ac_Oliva_'	Bajo	Bajo
'AC_oral', 'Ac_Girasol_'	'AC_oral'	'Ac_Girasol_'	Bajo	Bajo
'AC_oral', 'peras_peladas_mes'	'AC_oral'	'peras_peladas_mes'	Bajo	Bajo
'AC_oral', 'Ac_Girasol_'	'AC_oral'	'Ac_Girasol_'	Bajo	Bajo
'Frec_RX', 'manzanas_peladas_mes'	'Frec_RX'	'manzanas_peladas_mes'	Bajo	Bajo
'Frec_RX', 'Ac_Oliva_'	'Frec_RX'	'Ac_Oliva_'	Bajo	Bajo
'Frec_RX', 'peras_peladas_mes'	'Frec_RX'	'peras_peladas_mes'	Bajo	Bajo
'manzanas_peladas_mes','Ac_Oliva_'	'manzanas_peladas_mes'	'Ac_Oliva_	Bajo	Bajo
'manzanas_peladas_mes','Ac_Girasol_'	'manzanas_peladas_mes'	'Ac_Girasol_'	Bajo	Bajo
'manzanas_peladas_mes', 'peras_peladas_mes'	'manzanas_peladas_mes'	'peras_peladas_mes'	Bajo	Bajo
'Ac_Oliva_', 'Ac_Girasol_'	'Ac_Oliva_'	'Ac_Girasol_'	Bajo	Bajo
'Ac_Oliva_', 'peras_peladas_mes'	'Ac_Oliva_'	'peras_peladas_mes'	Bajo	Bajo
'Ac_Girasol_', 'peras_peladas_mes'	'Ac_Oliva_'	'peras_peladas_mes'	Bajo	Bajo

Tabla AIII 7

Diagnóstico C-95 (leucemia)				
Combinación de variables	**Var 1: nombre de la variable**	**Var 2: nombre de la variable**	**Var 1 valor**	**Var 2 valor**
'Estudios_diag', 'Sol_hs_mes'	Estudios_diag	Sol_hs_mes	No	Bajo
'Estudios_diag', 'carnesRojas_mes'	Estudios_diag	carnesRojas_mes	No	Bajo
'Estudios_diag', 'Galletas'	Estudios_diag	Galletas	No	Bajo
'Estudios_diag', 'carnesRojas_horno_mes'	Estudios_diag	carnesRojas_horno_mes	No	Bajo
'Estudios_diag', 'Legumbres'	Estudios_diag	Legumbres	No	Bajo
'Estudios_diag', 'AC_oral'	Estudios_diag	AC_oral	No	No
'Estudios_diag', 'Cigarrillos_acum'	Estudios_diag	Cigarrillos_acum	No	Bajo
'Sol_hs_mes', 'carnesRojas_mes'	Sol_hs_mes	carnesRojas_mes	Bajo	Bajo
'Sol_hs_mes', 'Legumbres'	Sol_hs_mes	Legumbres	Bajo	Bajo
'Sol_hs_mes', 'AC_oral'	Sol_hs_mes	AC_oral	Bajo	No
'Sol_hs_mes', 'Cigarrillos_acum'	Sol_hs_mes	Cigarrillos_acum	Bajo	Bajo
'carnesRojas_mes', 'Galletas'	carnesRojas_mes	Galletas	Bajo	Bajo
'carnesRojas_mes', 'carnesRojas_horno_mes'	carnesRojas_mes	carnesRojas_horno_mes	Bajo	Bajo
'carnesRojas_mes', 'Legumbres'	carnesRojas_mes	Legumbres	Bajo	Bajo
'carnesRojas_mes', 'AC_oral'	carnesRojas_mes	AC_oral	Bajo	No
'carnesRojas_mes', 'Cigarrillos_acum'	carnesRojas_mes	Cigarrillos_acum	Bajo	Bajo
'Galletas', 'carnesRojas_horno_mes'	Galletas	carnesRojas_horno_mes	Bajo	Bajo
'Galletas', 'Legumbres'	Galletas	Legumbres	Bajo	Bajo
'Galletas', 'AC_oral'	Galletas	AC_oral	Bajo	No

'Galletas', 'Cigarrillos_acum'	Galletas	Cigarrillos_acum	Bajo	Bajo
'carnesRojas_horno_mes', 'Legumbres'	carnesRojas_horno_mes	Legumbres	Bajo	Bajo
'carnesRojas_horno_mes','AC_oral'	carnesRojas_horno_mes	AC_oral	Bajo	No
'carnesRojas_horno_mes', 'Cigarrillos_acum'	carnesRojas_horno_mes	Cigarrillos_acum	Bajo	Bajo
'Legumbres', 'AC_oral'	Legumbres	AC_oral	Bajo	No
'Legumbres', 'Cigarrillos_acum'	Legumbres	Cigarrillos_acum	Bajo	Bajo
'AC_oral', 'Cigarrillos_acum'	AC_oral	Cigarrillos_acum	No	Bajo

Las tablas de los cánceres de colon, pulmón, piel, linfoma y leucemia muestran que el cruce de las variables seleccionadas por el algoritmo siempre da un valor bajo (ver gráficas en anexos).

No hubo resultados para los diagnósticos: C-50 (mama), C-55 (cuello de útero),C-61 (próstata).

Glosario

Aneuploidía: es una alteración en el número de cromosomas en una célula debido a pérdida o duplicación.

Asparagina (asn): es un aminoácido que puede encontrarse en múltiples alimentos. En la actualidad, está tomando mucha importancia su estudio puesto que es uno de los principales componentes en la reacción de Maillard que puede derivar en la producción de acrilamida, compuesto cancerígeno.

Clustering: del inglés *cluster*, que significa agrupar. Es decir, es el análisis de grupos (o agrupamiento), la tarea de agrupar objetos por similitud, en grupos o conjuntos de manera que los miembros del mismo grupo tengan características similares. Es la tarea principal de la minería de datos exploratoria y es una técnica común en el análisis de datos estadísticos.

Melanoma: es una enfermedad por la que se forman células malignas (cancerosas) en los melanocitos (células que dan color a la piel). Hay distintos tipos de cáncer que comienzan en la piel.

Mielógena: la leucemia mielógena aguda (LMA) es un cáncer que comienza dentro de la médula ósea. Esta es el tejido blando en el interior de los huesos que ayuda a formar las células sanguíneas. El cáncer crece a partir de las células que normalmente se convertirían en glóbulos blancos.

Mitosis: es el proceso por el cual una célula replica sus cromosomas y luego los secreta, produciendo dos núcleos idénticos durante la preparación para la

división celular. La mitosis generalmente es seguida por la división igual del contenido de la célula en dos células hijas que tienen genomas idénticos.

Mutágeno: sustancia y preparado que, por inhalación, ingestión o penetración cutánea, puede producir alteraciones genéticas hereditarias o aumentar su frecuencia.

Pirólisis: es una degradación térmica de una sustancia en ausencia de oxígeno, por lo que dichas sustancias se descomponen mediante calor, sin que se produzcan las reacciones de combustión.