UNIVERSIDAD MIGUEL HERNÁNDEZ

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

TRABAJO DE FIN DE GRADO EN ESTADÍSTICA EMPRESARIAL

PROBLEMAS DE OPTIMIZACIÓN PARA LA CLASIFICACIÓN: APLICACIÓN A LAS CIENCIAS DE LA SALUD.

Autor: Natalia Pérez Marín

Tutora: Mercedes Landete

Curso 2022/2023

1. Resumen

El cáncer colorrectal es una de las neoplasias más comunes en la actualidad y ha tenido una historia de reconocimiento y tratamiento que se ha desarrollado a lo largo del tiempo. Aunque los tumores de colon fueron descritos en épocas antiguas, el progreso significativo en el tratamiento y la comprensión de la enfermedad comenzó a tomar forma en el siglo XVII.

Actualmente, sigue siendo una enfermedad de alta incidencia, afectando a muchas personas en todo el mundo, incluyendo a la población española. Avances científicos y tecnológicos han permitido mejorar las opciones de tratamiento y brindar una mayor esperanza de supervivencia a los pacientes. Sin embargo, recibir un diagnóstico de cáncer es un desafío emocional y físico considerable para quienes lo enfrentan.

La noticia de un diagnóstico de cáncer colorrectal a menudo toma por sorpresa al paciente lo que puede generar un impacto significativo en su vida cotidiana. Los pacientes se encuentran ante la necesidad de reestructurar su rutina diaria, hacer ajustes en sus planes y enfrentar situaciones complejas relacionadas con el tratamiento, las operaciones y visitas médicas.

La variedad en los tratamientos es amplia: algunos pacientes pueden someterse a terapias cortas y efectivas, lo que les permite recuperar su vida normal con relativa rapidez. En otros casos, los tratamientos son más extensos y complejos, involucrando intervenciones quirúrgicas prolongadas y terapias más intensivas.

Cada paciente es único y su experiencia con el cáncer es personal y singular. Por esta razón, es esencial abordar a cada paciente con un enfoque individualizado y cuidadoso. El apoyo emocional, la educación sobre el diagnóstico y el tratamiento, y la atención médica integral son componentes vitales en la atención de los pacientes con cáncer colorrectal. A medida que la ciencia y la medicina continúan avanzando, la esperanza y las oportunidades para una supervivencia plena siguen creciendo, brindando un mensaje de esperanza a quienes enfrentan esta enfermedad.

1.1 ¿Qué es el cáncer colorrectal?

Hablamos de este cáncer cuando el tumor maligno se encuentra situado en el colon o en el recto. Para poder asegurar la aparición de cáncer es necesario que transcurran varios

años desde que se transforma la primera célula en un pólipo hasta que comienza la sintomatología.

Aunque solemos hablar de cáncer colorrectal, engloba a dos tumores diferentes: *cáncer de colon y cáncer de recto.*

Un pólipo es un crecimiento anormal en la mucosa del colon o el recto, y pueden ser benignos (no cancerosos) o precursores de cáncer (adenomas). Es importante señalar que la mayoría de los cánceres colorrectales se desarrollan a partir de pólipos adenomatosos. Cuanto mayor es el tamaño que va cogiendo el pólipo mayor acumulación de mutaciones genéticas. La detección temprana y la eliminación de pólipos es fundamental para reducir el riesgo de desarrollar este tipo de cáncer.

Por eso, los exámenes de detección periódicos son esenciales en la población de más de 50 años. Gracias al Programa de Detección Precoz de Cáncer Colorrectal iniciada en 2016, es un programa dirigido a hombres y mujeres de entre 50 y 69 años, reciben una carta en la que incluye un kit para la realización de un análisis de heces que posteriormente han de llevar a su centro médico que será analizado, si el resultado es positivo el facultativo se pone en contacto para avanzar otro paso más y realizar una colonoscopia. La finalidad de este cribado es prevenir y poder tratarlo antes de que aumente su tamaño.

1.2 Síntomas del cáncer colorrectal

- Sangre en las heces: Es uno de los síntomas más comunes. Puede manifestarse de dos maneras principales: sangre roja en las heces o sangre negra que se mezcla con las heces y causa deposiciones de color negro, conocidas como melenas.

Cuando el sangrado persiste durante un período de tiempo y no se detecta ni se busca atención médica para un diagnóstico y tratamiento, puede llevar a la anemia. La anemia puede ser resultante con síntomas como: sensación de falta de aire, cansancio y fatiga, palpitaciones y mareo.

La detección temprana y el tratamiento oportuno son cruciales para mejorar las perspectivas de recuperación y reducir la gravedad de los síntomas.

La población mayor de 50 años es llamada para la realización de una analítica de heces con la finalidad de prevenir un posible tumor y en caso de tenerlo poder tratarlo cuanto antes.

- Cambio en el ritmo de las deposiciones: Puede aparecer diarrea o estreñimiento en personas con un ritmo intestinal normal.
- Dolor abdominal

- Pérdida de peso sin causa aparente

1.3 Tratamiento.

Una vez confirmado el diagnóstico, surge el objetivo crucial del tratamiento, que radica en eliminar de manera exhaustiva y, en la medida de lo posible, el cáncer mismo. Este logro puede ser alcanzado mediante distintos enfoques terapéuticos, que comprende desde la aplicación de un único tratamiento hasta todas las intervenciones médicas y técnicas, tales como la cirugía, radioterapia, quimioterapia y en nuestra ocasión, terapias diseñadas específicamente para combatir con un tumor determinado y complicado de eliminar.

1.4 Herramientas estadísticas para el análisis de datos de pacientes con cáncer colorrectal

Todo el análisis se ha realizado con R-Studio, una interfaz basada en lenguaje R.

R presenta diversas ventajas en contraste con otras herramientas de análisis estadístico. En primer lugar, se trata de una herramienta de código abierto, lo que implica la creación de nuevas aplicaciones o alterar las ya existentes. Además, para poder trabajar cómodamente con él no tiene ningún costo adicional, lo que implica una ventaja más para el programa.

Por otro lado, R es conocido por las numerosas librerías disponibles las cuales nos permiten visualizar gráficos y analizar bases de datos complejas. En este proyecto se han utilizado las siguientes librerías:

- Tidyverse: se trata de una colección de paquetes dentro de R adecuado para el análisis y manipulación de datos, esta librería contiene varios paquetes que han sido de gran ayuda para un correcto desarrollo:

- Ggplot2: se utiliza para la creación de gráficos en R de alta calidad. Ha sido utilizada para la creación del gráfico de SVM para poder ver con claridad los pacientes que se encuentran mal clasificados.

- Dplyr: Paquete de gran utilidad para la limpieza, organización, transformación y resumen de datos. Es utilizado para el procesamiento y depuración de datos para la creación de un análisis descriptivo.

- e1071: Este paquete se usa para implementar algoritmos, en especial en el ámbito de la minería de datos. Se ha puesto en funcionamiento para la creación de Support Vector Machines (SVM).

Crear un proyecto con una base de datos de pacientes oncológicos donde todas las variables y valores de éstas son reales puede ofrecer a la comunidad científica una amplia gama de beneficios y contribuciones en el campo de la medicina, investigación y atención médica. Puede servir de ayuda a la investigación clínica y científica con la identificación de patrones de enfermedades, eficacia de tratamientos y tendencias en la salud. Por otro lado, al desarrollo y evaluación de tratamientos consiguiendo que algunos fármacos que se utilizan dejen de hacer con el fin de evitar mayor cantidad de efectos secundarios.

2. Introducción al problema

Actualmente, hospitales y centros de investigación se han unido para estudiar y mejorar el tratamiento de pacientes con tumores malignos. En nuestro caso, los pacientes con tumores cancerosos se espera que sea uno de los grupos más numerosos, con alrededor de 42.721 nuevos casos durante el año 2023.

Nuestra base de datos contiene información sobre pacientes que previamente han recibido tratamientos que no han sido efectivos y que ahora han accedido a una terapia exclusiva. El primer paso es detectar el tumor y determinar si es maligno o benigno. En caso de ser maligno, se inicia la fase de limpieza mediante una operación quirúrgica. Durante esta intervención, se administra al paciente una mezcla de líquido y quimioterapia a través de la barriga. El proceso se controla por la monitorización de la temperatura y las áreas donde se expande el líquido administrado.

La base de datos recopila una serie de variables relevantes para el estudio. Se registran características básicas como el sexo, fecha de nacimiento y la edad de los pacientes. Sin embargo, la variable con mayor relevancia para la investigación es el fármaco administrado, diferenciado entre “Paclitaxel” y “Mitomicina”, dos tipos de antibióticos que se utilizan para ralentizar y/o detener el crecimiento de células cancerosas en el cuerpo.

Además, se registran datos como el diámetro de la cintura, altura, peso e índice de masa corporal (IMC) de cada uno de los pacientes que son aceptados en este tratamiento exclusivo. También se analizan las concentraciones de plasma y líquido en diferentes momentos del proceso, así como las temperaturas del líquido administrado en diferentes zonas de la pelvis. Por último, se registran los volúmenes teóricos e instalados, éstos representan la cantidad de líquido que se puede administrar (volumen teórico) y

finalmente la cantidad administrada en función del médico y situación interna del paciente (volumen instalado).

Este fichero proporciona información valiosa para el estudio y la mejora de los tratamientos en pacientes oncológicos con tumores malignos. El análisis de las diferentes características permite evaluar la eficacia de la terapia exclusiva y comprender cómo el líquido y la quimioterapia afectan a los pacientes en términos de concentración, temperatura y expansión en distintas áreas de la pelvis. Esto puede ayudar a los profesionales de la salud a tomar decisiones más informadas y a desarrollar estrategias de tratamiento más efectivas en el futuro.

Palabras clave: Mitomicina; Paclitaxel; Tumor; Tumores oncológicos; Pacientes; Conc_plasma_45; Conc_plasma_60; Support Vector Machine; SVM

3. Estudio de variables

La base de datos proporciona información sobre características básicas de los pacientes, como su sexo, fecha de nacimiento y edad a la que le ha sido detectado. Sin embargo, la variable principal de interés en este estudio es el fármaco administrado que se diferencia en dos tipos: Paclitaxel y Mitomicina. Ambos son antibióticos con la capacidad de retardar o detener el crecimiento de células cancerosas en el cuerpo.

Otro aspecto de importancia es el concepto de "Día HIPEC". Este término se refiere al día en que se administra unos de los dos fármacos mencionados anteriormente, y también es el momento en que se recogen muestras para analizar las variables que se estudiarán posteriormente. Esto implica que se lleva a cabo un seguimiento de los pacientes después de recibir el tratamiento con el fármaco seleccionado, y se recopilan datos específicos para su posterior análisis.

Es importante destacar que la diferenciación entre "Paclitaxel" y "Mitomicina" se debe a sus propiedades farmacológicas y a sus efectos específicos en las células cancerosas. Ambos fármacos se utilizan con el objetivo de controlar y combatir el crecimiento maligno de las células propias de tumores oncológicos.

En resumen, esta base de datos recopila información demográfica básica de los pacientes,

junto con detalles cruciales sobre el fármaco administrado. La distinción entre "Paclitaxel" y "Mitomicina" es esencial para el estudio, ya que estos medicamentos tienen un impacto significativo en la detención o ralentización del crecimiento de estas células cancerosas.

Podemos concluir con que esta base de datos proporciona una base sólida para investigadores y estudios en el campo del tratamiento del cáncer, permitiendo un mejor entendimiento de los efectos de estos fármacos en los pacientes y su eficacia en la lucha contra la enfermedad.

- SC (m2): Diámetro de cintura
- Talla (m): Altura del paciente en estudio
- Peso (Kg): Peso por paciente
- IMC (Kg/m2): Índice de Masa Corporal
- Con_plasma_15/30/45/60: Concentración de plasma en el paciente a los 15, 30, 45 y 60 minutos.
- Con_Liquido_peri_pelvis 15/30/45/60: Concentración de líquido en la pelvis en la zona más baja.
- Con_Liquido_peri_angulo 15/30/45/60: Concentración de líquido en la zona anterior a la pelvis y controlado cada quince minutos.
- Con_Liquido_peri_espacio 15/30/45/60: Concentración de líquido en la zona anterior de la pelvis cerca de la variable "peri_angulo" que los médicos diferencian como "espacio".
- Temp_liquido_peri_pelvis 15/30/45/60: Temperatura a la que se encuentra el fármaco una vez inyectado en los diferentes tiempos en la zona más baja de la pelvis.
- Temp_liquido_peri_angulo 15/30/45/60: Temperatura del líquido administrado cada quince minutos.
- Temp_liquido_peri_espacio 15/30/45/60: Temperatura del líquido administrado cada quince minutos.
- Volumen Teórico: Magnitud en litros calculada por perímetro de la barriga e índice de masa corporal inyectable a cada paciente.,
- Volumen Instalado: Cantidad en litros administrada una vez entrado el paciente a quirófano y admitido por su barriga.

4. Análisis descriptivo

El análisis descriptivo llevado a cabo se relaciona con los pacientes que han sido sometidos a un tratamiento con tumores oncológicos, tras haber sido sometido a un riguroso proceso de preprocesamiento y depuración, arroja luz sobre los resultados esenciales que sientan las bases para la siguiente fase de investigación. En esta segunda etapa, se enfoca en el estudio minucioso de las solicitudes de utilización de los diferentes materiales y variados enfoques terapéuticos.

La meticulosa presentación de los datos descriptivos no solo proporciona una visión integral, sino que también posibilita una comprensión más profunda y detallada de la distribución y evolución de los tumores en relación con los tratamientos empleados. A través de esta exposición, se pone de relieve, en primer plano, el recuento total de observaciones (n), la media descrita para cada una de las características de estudio, la asimetría que muestra si los datos están sesgados hacia un lado, tener valores NA sugiere que no hay suficiente variación en los datos para poder calcularla correctamente. Por otro lado, la curtosis mide la forma de distribución, teniendo en cuenta si existe suficiente variabilidad. Error estándar de la Media (se) nos indica con mayor precisión la estimación de la media, cuánto mayor es el valor de error menos correcta será la media, se sugiere que el error sea lo más próximo a cero, de esta forma, la media para cada variable será más correcta.

Por último, obtener valores infinitos en las columnas de mínimo, máximo y rango, nos indica que hay valores en la columna excesivamente pequeños o grandes encontrándose fuera del rango de cálculo para estadísticas descriptivas. Estos casos se pueden tratar como valores atípicos (outliers), en algunas ocasiones podrían eliminarse del estudio lo que implicaría no mostrar unos resultados totalmente reales y válidos.

Tabla 1 - Análisis descriptivo

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
SC (m2)	2	49	1.5842857	0.6281388	1.70000	1.6812195	0.2075640	0.0000	2.21000	2.21000	-1.81331450	2.0795080	0.08973411
Talla (m)	3	49	1.4318367	0.5458811	1.62000	1.5378049	0.1037820	0.0000	1.80000	1.80000	-2.14495353	2.8291520	0.07798301
Peso (kg)	4	49	66.0673469	28.9497653	68.00000	68.8365854	17.7912000	0.0000	107.00000	107.00000	-1.12205344	0.7160817	4.13568076
IMC (kg/m2)	5	43	28.1726886	4.8475726	27.88762	28.0738534	5.5594233	20.9042	36.57979	15.67559	0.19595994	-1.2738608	0.73924790
Conc_plasma_15	6	49	1.0237469	1.3199759	0.12070	0.8685341	0.1666442	0.0000	4.46780	4.46780	0.83702579	-0.8168843	0.18856798
Conc_plasma_30	7	49	0.9815980	1.2508502	0.09910	0.8425585	0.1469257	0.0000	4.31450	4.31450	0.83073428	-0.8089185	0.17869289
Conc_plasma_45	8	49	0.9647816	1.2409896	0.10420	0.8250463	0.1544869	0.0000	4.19930	4.19930	0.83814544	-0.8321709	0.17728423
Conc_plasma_60	9	49	0.9227102	1.1789447	0.11360	0.7873683	0.1624930	0.0000	3.96190	3.96190	0.82081595	-0.8691297	0.16842067
Conc_liquido_peri_pelvis_15	10	49	17.1046714	10.9690137	18.30370	16.8912171	12.4218158	0.0000	42.10400	42.10400	0.13326697	-0.9839724	1.56700195
Conc_liquido_peri_pelvis_30	11	49	16.6749490	9.8170083	15.92590	16.6611829	10.8889557	0.0000	41.18120	41.18120	0.11752808	-0.7442436	1.40242976
Conc_liquido_peri_pelvis_45	12	49	15.1907837	9.4343363	14.79150	15.1894244	10.9199420	0.0000	33.35190	33.35190	0.01033884	-1.1967846	1.34776233
Conc_liquido_peri_pelvis_60	13	49	14.4108898	9.4902801	14.70470	14.1581878	11.0057846	0.0000	36.97080	36.97080	0.23981226	-0.9394517	1.35575430
Conc_liquido_peri_angulo_15	14	49	14.6230184	11.0851864	11.39660	13.8932390	15.5441714	0.0000	42.16540	42.16540	0.41930968	-0.6616451	1.58359805
Conc_liquido_peri_angulo_30	15	49	14.3069122	9.8399214	12.39730	14.0223659	13.1856514	0.0000	37.56370	37.56370	0.14020302	-0.9441315	1.40570305
Conc_liquido_peri_angulo_45	16	49	12.6276347	10.6026321	10.59320	11.7840854	13.0944715	0.0000	37.27290	37.27290	0.54236575	-0.6916251	1.51466172
Conc_liquido_peri_angulo_60	17	49	11.4444449	9.8810747	9.08840	10.7347854	13.4744618	0.0000	37.54050	37.54050	0.54159007	-0.6765596	1.41158210
Conc_liquido_peri_espacio_15	18	49	15.3692735	11.3596931	12.39100	14.9166439	14.6476432	0.0000	46.47850	46.47850	0.31510299	-0.6622642	1.62281329
Conc_liquido_peri_espacio_30	19	49	14.9574286	10.4504893	13.23700	14.6541829	12.3774861	0.0000	44.57890	44.57890	0.25868428	-0.4901071	1.49292704
Conc_liquido_peri_espacio_45	20	49	14.0400327	9.9879283	12.60710	13.7041488	11.2576783	0.0000	39.07810	39.07810	0.23392122	-0.8473075	1.42684690
Conc_liquido_peri_espacio_60	21	49	14.5639735	10.5296818	14.69510	14.0187024	11.2406284	0.0000	49.91100	49.91100	0.66980222	0.6223653	1.50424025
Temp_liquido_peri_pelvis_15	22	49	40.4244898	8.5049135	42.30000	42.1878049	1.0378200	0.0000	43.60000	43.60000	-4.36734764	17.8369567	1.21498765
Temp_liquido_peri_pelvis_30	23	49	42.3000000	1.0946841	42.70000	42.4585366	0.7413000	38.5000	43.60000	5.10000	-1.51059396	1.9997686	0.15638345
Temp_liquido_peri_pelvis_45	24	49	40.5489796	8.5119309	42.60000	42.3243902	0.8895600	0.0000	43.60000	43.60000	-4.39884462	18.0243918	1.21599013
Temp_liquido_peri_pelvis_60	25	49	40.4938776	8.5011374	42.50000	42.2487805	0.7413000	0.0000	43.60000	43.60000	-4.39733033	18.0168570	1.21444820
Temp_liquido_peri_angulo_15	26	49	40.9632653	1.2047984	40.90000	40.9682927	1.4826000	38.9000	43.20000	4.30000	0.11210847	-1.2272293	0.17211405
Temp_liquido_peri_angulo_30	27	49	41.0653061	1.0823491	40.90000	41.1000000	1.0378200	39.0000	42.70000	3.70000	-0.02492111	-0.9034923	0.15462130
Temp_liquido_peri_angulo_45	28	49	41.1020408	1.0619341	40.90000	41.1170732	1.0378200	39.0000	42.90000	3.90000	0.13598528	-1.1165099	0.15170487
Temp_liquido_peri_angulo_60	29	49	41.1204082	1.0770290	40.90000	41.1439024	1.3343400	39.0000	43.00000	4.00000	-0.04484590	-1.0020582	0.15386129
Temp_liquido_peri_espacio_15	30	49	41.2897959	1.1996953	41.50000	41.3536585	1.4826000	38.7000	43.00000	4.30000	-0.41013032	-0.9587584	0.17138504
Temp_liquido_peri_espacio_30	31	49	40.4959184	6.0084302	41.50000	41.3609756	1.3343400	0.0000	43.00000	43.00000	-6.23447048	39.1171604	0.85834717
Temp_liquido_peri_espacio_45	32	49	40.4448980	5.9900981	41.50000	41.2902439	1.0378200	0.0000	43.00000	43.00000	-6.26950061	39.4199386	0.85572830
Temp_liquido_peri_espacio_60	33	49	41.3795918	1.1454583	41.60000	41.4317073	1.3343400	38.9000	43.00000	4.10000	-0.40842329	-0.8967615	0.16363691
VOLUMEN TEÓRICO	34	49	3342.4489796	939.6775391	3480.00000	3502.4390244	296.5200000	0.0000	4420.00000	4420.00000	-2.59674610	6.9588666	134.23964845
VOLUMEN INSTILADO	35	49	3626.7142857	1208.0030353	3700.00000	3693.4146341	444.7800000	0.0000	6130.00000	6130.00000	-1.06445691	3.2042697	172.57186219

Fuente: Elaboración propia

5. Support Vector Machine

El SVM (Máquinas de Vectores de Soporte) fue presentado en 1995 por *Cortes y Vapnik*. Se basa en la estrategia de mantener fijo el valor del riesgo empírico minimizando el intervalo de confianza. El SVM construye un hiperplano de separación óptimo para las dos clases diferenciadas, *paclitaxel* y *mitomicina*. Si los datos son linealmente separables, el margen se trata como "duro"; si no, el margen es "suave". El SVM ha demostrado ser una herramienta muy efectiva para la clasificación supervisada, su objetivo es maximizar el margen entre dos hiperplanos paralelos equidistantes y minimizar el número de vectores mal clasificados. Este proceso es uno de los más utilizados para conseguir la reducción de variables y poder obtener una solución firme y válida.

Por otro lado, la *obtención de la Frontera de Pareto de SVM con Selección de Características* se presenta como una herramienta metaheurística. La selección de características en un modelo SVM se lleva a cabo mediante variables binarias donde se

indica qué variables están seleccionadas, para el estudio utilizaremos "*-1"* para indicar que el fármaco administrado es *paclitaxel* y "*1" mitomicina.*

Un modelo SVM con selección de características se convierte en un modelo no convexo, por lo que, para poder obtener resultados fiables es necesario emplear técnicas propias de problemas no convexos.

Centrándonos en el modelo de SVM, sirve para categorizar vectores representados por un par $(X_i, Y_i) \in R^n x \{-1, 1\}$, donde *n* es el número de características que tiene cada paciente (Peso, Con_Liquido_peri_espacio_15, etc.), X_i integra las características del vector *i e* Y_i indica a cuál de las dos clases de Ω puede corresponder el vector *i*, sabiendo que diferenciamos entre dos fármacos; *mitomicina y paclitaxel*.

Si las dos clases acaban totalmente diferenciadas en el hiperplano, podemos concluir con que, Ω es linealmente separable, por lo tanto la existencia de valores para $v \in R^n$, $\theta \in R$ y $\mu \in R+0$, afirmando que todos los vectores de la clase $Yi=1$ cumplirán la restricción $v^T * x_i \leq \theta - \mu$ y los vectores con $Y_i = -1$ satisfacen la restricción $v^T * x_i \geq \theta + \mu$. Al hablar de diferenciación del hiperplano entre los conjuntos se refiere a que la función $f(x) = w^T * x + b$ es capaz de separarlos de forma óptima maximizando la distancia entre dos hiperplanos paralelos y minimizando el total de errores de clasificación. Para alcanzar este propósito, el modelo de Support Vector Machine de margen duro minimiza el equilibrio entre los dos objetivos, denominados, riesgo estructural y riesgo empírico.

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \qquad (1)$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i \qquad i = 1, \ldots, m, \qquad (2)$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, m, \qquad (3)$$

Figura 01 - Formulación SVM

Fuente: Support Vector Machine with feature selection: A multiobjective approach

El modelo utiliza dos objetivos; $O_1 = \frac{1}{2}||w||^2$ y $O_2 = \sum_{i=1}^{m} \xi_i$, que representan el riesgo empírico y estructural, respectivamente, y pueden optimizarse de forma separada. El modelo se utiliza para clasificar dos objetivos de clases mediante la definición de hiperplanos paralelos, cuyos coeficientes vienen dados por el vector *w*, este vector tendrá tantos coeficientes como variables tenga nuestra base de datos, exactamente, treinta y cuatro. El parámetro *C* controla que no se clasifiquen de forma incorrecta los datos.

Las restricciones (2) y (3) aseguran que, los vectores $Y_i = 1$ o $Y_i = -1$ satisfacen $Y_i(w^T x_i + b) \geq 1$ y $Y_i(w^T x_i + b) \leq 1$, respectivamente.

6. Análisis del problema.

Para conseguir una solución con SVM en RStudio es necesario trabajar con la librería QuadProg, pero en ésta no se podían añadir variables que no fuesen cuadráticas, es decir, para llevar a cabo el problema con dicha librería tenían que ser semidefinidas positivas. Ante este inconveniente, nos adaptamos al fallo e implementamos el mismo problema, pero con la librería lpSolve propia de rstudio y que la diferencia en que las "w" que en SVM están al cuadrado ahora podrán ser lineales, esta variación no se verá afectada en el análisis de resultados al igual que tampoco lo hará en su desarrollo.

En primer lugar, escribimos el problema a mano con la finalidad de poder imaginarlo antes de comenzar a trabajar en R.

$$Min \;\; W_1 + W_2 + W_3 + \ldots + W_{34} + \xi_1 + \xi_2 + \xi_3 + \ldots + \xi_{49}$$

$$\begin{aligned}
\text{s. a} \quad & W_1 * X_1 + W_1 * X_2 + \ldots + W_1 * X_{49} + b \geq 1 - \xi_1 \\
& W_2 * X_1 + W_2 * X_2 + \ldots + W_2 * X_{49} + b \geq 1 - \xi_2 \\
& W_3 * X_1 + W_3 * X_2 + \ldots + W_3 * X_{49} + b \geq 1 - \xi_3 \\
& \ldots \\
& \ldots \\
& W_{34} * X_1 + W_{34} * X_2 + \ldots + W_{34} * X_{49} + b \geq 1 - \xi_{49}
\end{aligned}$$

$$W_1 , \quad W_2 , \quad W_3 , \quad \ldots , \quad W_{34} \geq 0$$

En el análisis de restricciones, se examinan todas las variables de forma individual para cada paciente. Esto implica considerar diferentes factores y características relacionadas con el tratamiento y la condición médica del paciente. Si el fármaco administrado es Paclitaxel, se multiplica toda la restricción por -1 como parte del proceso de ajuste y cálculo. Por otro lado, si se utiliza Mitomicina, se realiza una multiplicación por uno. Esto se hace para tener en cuenta los efectos específicos de cada fármaco en el contexto del problema en cuestión.

Además de las restricciones, también se agrega una nueva columna que representa la influencia del fármaco en estudio. Esta columna puede tomar valores de 1 o -1, según el fármaco administrado al paciente. Esto permite considerar y cuantificar cómo afecta el fármaco a la solución óptima del problema.

Por último, se utiliza el concepto de épsilon para formar un vector. Este vector se compone de unos y ceros, donde se establece un uno en la posición correspondiente a la restricción activa y ceros en las demás posiciones. Con esto, se crea una matriz identidad cuadrada de cuarenta y nueve filas y cuarenta y nueve columnas. Cada fila y columna representan una restricción, y la diagonal de la matriz tiene valores de unos. Este enfoque permite establecer de manera eficiente las restricciones y sus interacciones en el problema.

Al considerar un valor de épsilon igual a 0.5^1 de precisión, se puede concluir que las características "conc_plasma_45" y "conc_plasma_60" son las únicas que influyen significativamente en este problema. Estas características tienen valores específicos, con "conc_plasma_45" con un valor de 0.3747 y "conc_plasma_60" con un valor de 0.0002624. Estos valores indican la importancia relativa de cada característica en la solución óptima.

Al analizar los valores de épsilon, se encuentra que la mayoría de ellos se mantienen dentro del hiperplano óptimo correspondiente al problema. Específicamente, un 63.2653% de los valores de épsilon son iguales a cero, lo que implica que estas restricciones se encuentran en el hiperplano óptimo. Esto sugiere que estas restricciones tienen un impacto relevante en la solución final.

En resumen, el análisis de las restricciones y la influencia de los fármacos en el problema permite comprender mejor cómo cada variable y característica afecta la solución óptima. La consideración de los valores de épsilon y su relación con el hiperplano óptimo

proporciona información valiosa para tomar decisiones informadas y optimizar el proceso de tratamiento médico.

Se procede a cambiar el valor de épsilon con el fin de observar modificaciones en las características e influencia de estas, así como el valor de épsilon y el porcentaje de clasificación correcta. En este caso, se realizará un cambio multiplicando el valor de épsilon por 1.5. Al realizar este cambio, se obtienen las mismas características influyentes que en el caso anterior, así como el mismo porcentaje de épsilon mal clasificados.

1) este valor puede verse modificado según las necesidades del usuario, teniendo en cuenta que a mayores valores de épsilon su nivel de buscar características mal clasificadas aumentará

Sin embargo, se observa una diferencia en los valores de las características. En particular, el valor de la característica "conc_plasma_45" aumenta a 0.450826, mientras que el valor de la característica "conc_plasma_60" disminuye a 0.0001860. Este resultado sugiere que existen dos características que son de gran importancia a la hora de analizar el problema. Además, se puede deducir que, al aumentar el valor de épsilon, solo se verán afectados los valores de estas características, pero no su influencia en la solución del problema.

Este hallazgo refuerza la idea de que las características "conc_plasma_45" y "conc_plasma_60" desempeñan un papel fundamental en la resolución óptima del problema. Sus valores modificados indican la relevancia relativa de cada una de ellas en la toma de decisiones y en la solución final.

Es importante destacar que, a pesar del cambio en el valor de épsilon, el porcentaje de mal clasificados se mantiene exactamente igual que en el caso anterior. Esto sugiere que el ajuste en el valor de épsilon no tiene un impacto significativo en la clasificación correcta de los datos.

Por otro lado, se vuelve a modificar el valor de épsilon con el fin de encontrar una mejor solución al problema. En este caso, modificamos épsilon igual a uno. Tras realizar los cambios necesarios seguimos manteniendo los mismos resultados que cuando épsilon valía 0.5.

Al considerar otra modificación de épsilon, esta vez tomando el valor de 0.05, se pueden observar nuevos resultados en relación con las características y la clasificación de los

datos. Al disminuir el valor de épsilon, se introduce una característica adicional que se suma a las anteriormente mencionadas, conc_plasma_30 es la nueva característica influyente unida a conc_plasma_45 y conc_plasma_60. Estas características toman los valores de 0.016106071, 0.034573662 y 0.000587706 respectivamente.

En base a estos resultados, podemos afirmar que la concentración de plasma dentro del paciente después de administrarle la dosis correspondiente de medicamento y tomada la temperatura en los diferentes tiempos sigue siendo la característica más influyente dentro del estudio. Los valores obtenidos para esta característica en particular indican su relevancia en el proceso de clasificación de los datos.

Estos resultados muestran que existe margen para mejorar el proceso de clasificación y ajustar los parámetros utilizados en el modelo. La elección de épsilon, en este caso, tiene un impacto significativo en los resultados obtenidos. Es posible que al seguir probando diferentes valores de épsilon se pueda lograr una mejor clasificación de los datos, reduciendo así el porcentaje de clasificaciones incorrectas.

Es importante considerar que los resultados y conclusiones mencionados anteriormente se basan en los datos y características específicas utilizadas en este estudio. Dependiendo del contexto y los datos disponibles, los resultados pueden verse afectados y variar. Por lo tanto, es esencial realizar análisis exhaustivos y experimentos adicionales para obtener conclusiones más sólidas y generalizables.

Como conclusión, al modificar épsilon por un valor menor del que inicialmente se propuso, se introduce una característica adicional a las existentes con las anteriormente estudiadas, confirmando así que la concentración de plasma tras administrar el fármaco sigue siendo la más influyente dentro del estudio. Sin embargo, se observa una alta tasa de clasificación incorrecta, al igual que ocurre con los valores de épsilon anteriores, lo que nos indica la necesidad de ajustar los parámetros del modelo y explorar otras opciones para mejorar la precisión de la clasificación de datos.

Con los resultados que se han obtenido y analizando que la variable más importante es concentración de plasma, se estudia cómo se distribuyen los datos iniciales con el fin de poder obtener conclusiones con mayor contundencia.

Comenzamos por la medición de concentración de plasma a los 45 minutos para cada paciente teniendo en cuenta el fármaco que se le ha administrado.

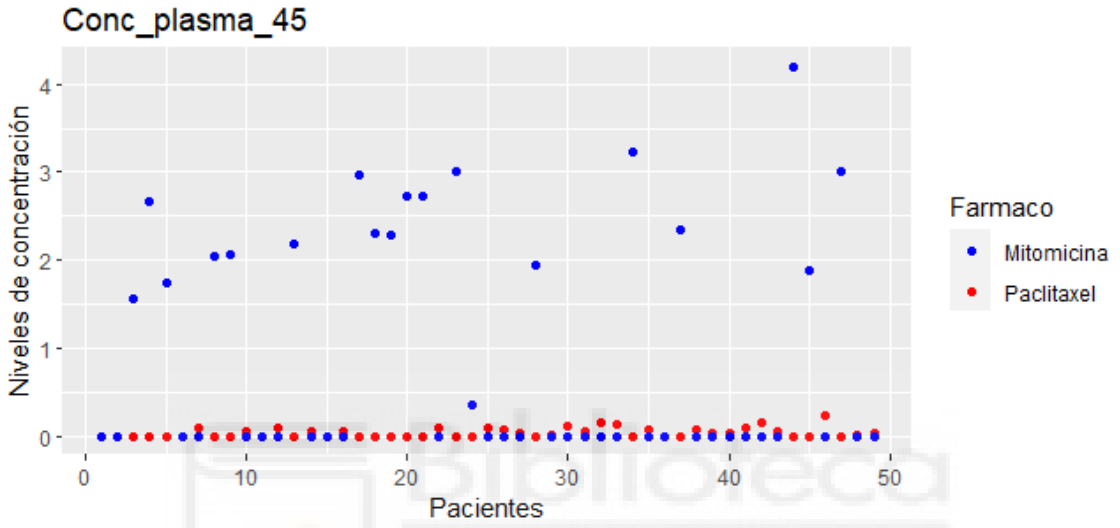

Figura 02 - Niveles de concentración de plasma a 45 minutos de administrarlo

Fuente: Elaboración propia

El análisis de la nube de puntos proporciona información relevante sobre la administración y efectos de los fármacos “paclitaxel” y “mitomicina” en pacientes con tumores malignos. Al observar el gráfico, se evidencia que el fármaco con mayor administración por parte de los médicos es “paclitaxel”, utilizado en el tratamiento de pacientes con tumores cancerosos.

Veintinueve pacientes han sido tratados con este fármaco específico.

Sin embargo, es notable que los valores de concentración de “paclitaxel” están muy próximos a cero en el gráfico, lo que sugiere que, a pesar de ser administrado en un número significativo de pacientes, no parece estar alcanzando una concentración terapéutica adecuada. Esto podría ser motivo de preocupación, ya que para que el fármaco sea efectivo, es importante que alcance una concentración suficiente en el cuerpo para detener la división celular y prevenir la formación de nuevas células.

Por otro lado, el fármaco “mitomicina” también ha sido utilizado en el tratamiento de pacientes con tumores malignos. Sus datos en el gráfico están dispersos por todo el rango de concentraciones, lo que indica que su administración ha resultado en diferentes niveles de concentración en diferentes pacientes. Aunque una gran parte de los valores también están próximos a cero, lo que podría indicar la necesidad de ajustar la dosis o el régimen de administración para lograr una concentración más efectiva.

Es esencial tener en cuenta que, en el análisis del gráfico, cuando el fármaco administrado no era el que se estaba analizando en el paciente, se asignó un valor de cero para mantener el tamaño de la muestra en cuarenta y nueve pacientes. Esta práctica ayuda a mantener la integridad de los datos y asegurar que los resultados sean comparables y significativos.

En cuanto a las características de los fármacos, el *paclitaxel* es un inhibidor de la mitosis que pertenece a la subcategoría de los taxanos, derivados de productos naturales de origen vegetal. Su mecanismo de acción busca detener la división celular y prevenir la formación de nuevas células. Es importante tener un control preciso de la cantidad administrada, ya que el exceso puede causar daños al sistema nervioso.

Por otro lado, la *mitomicina* pertenece al grupo de los antibióticos antitumorales. Su objetivo es cambiar el ADN de las células cancerosas para impedir que crezcan y se multipliquen. Este enfoque terapéutico busca atacar directamente el material genético de las células malignas y limitar su capacidad de proliferación.

En conclusión, el análisis de la nube de puntos ha revelado que el *paclitaxel* es el fármaco más administrado por los médicos en el tratamiento de tumores malignos, aunque sus concentraciones no parecen ser adecuadas en muchos casos. Por otro lado, la *mitomicina* muestra una mayor variabilidad en sus concentraciones, pero también se encuentra presente en una parte significativa de los pacientes. Estos hallazgos pueden ser valiosos para mejorar el tratamiento de pacientes oncológicos y optimizar la administración de estos fármacos para lograr mejores resultados en la lucha contra el cáncer.

El análisis de la nube de puntos para la concentración de plasma a los sesenta minutos proporciona información adicional sobre la efectividad y distribución de los fármacos *paclitaxel* y mitomicina en el tratamiento de pacientes con tumores malignos.

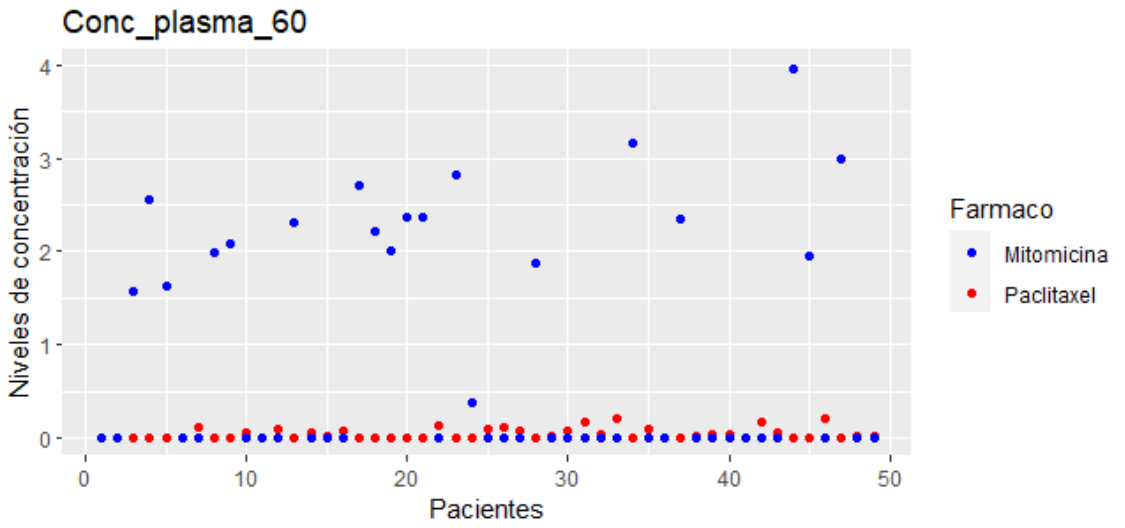

Figura 03 - Niveles de concentración de plasma a 60 minutos de administrarlo

Fuente: Elaboración propia

En comparación con el gráfico de dispersión de concentración de plasma a los cuarenta y cinco minutos, el gráfico para los sesenta minutos muestra una similitud visual significativa. Los niveles de concentración para el fármaco *paclitaxel* en el rango de pacientes de 30 a 50 son más elevados que a los cuarenta y cinco minutos, aunque siguen siendo muy similares en general.

Esta información sugiere que, en algunos pacientes, la concentración del fármaco *paclitaxel* aumenta ligeramente a los sesenta minutos después de la administración.

Por otro lado, el fármaco *mitomicina* sigue exhibiendo una distribución dispersa en todo el gráfico, lo que indica que la concentración de este fármaco varía considerablemente entre diferentes pacientes a los sesenta minutos. Sin embargo, es importante destacar que en algunos casos, como en los pacientes 20 y 21, los niveles de concentración son muy cercanos, lo que sugiere que la respuesta al fármaco puede ser similar en estos casos específicos.

Estos hallazgos son relevantes para el estudio y mejora de los tratamientos en pacientes oncológicos con tumores malignos. La observación de la concentración de plasma a diferentes intervalos de tiempo, como los cuarenta y cinco y sesenta minutos, proporciona información valiosa sobre cómo los fármacos *paclitaxel* y *mitomicina* se distribuyen y se comportan en el cuerpo de los pacientes, aun siendo de diferente sexo la reacción del fármaco puede ser muy similar.

La información visual obtenida de estos gráficos de dispersión permite evaluar la eficacia y consistencia del tratamiento con estos fármacos. Identificar patrones de concentración en diferentes momentos del proceso puede ayudar a ajustar la dosis y los tiempos de administración para lograr un efecto terapéutico óptimo.

Además, el hecho de que algunos pacientes muestran concentraciones de fármaco muy cercanas entre sí, como se observa en el caso de los pacientes 20 y 21 con *mitomicina*, puede requerir una atención especial en la monitorización y ajuste de la terapia para maximizar los resultados y minimizar efectos secundarios.

Con el objetivo de comparar los niveles de concentración para todos los pacientes diferenciando el fármaco administrado y sin tener en cuenta si la muestra está tomada a los cuarenta y cinco o sesenta minutos se trata de evaluar si las concentraciones de ambos fármacos son similares o si hay diferencias significativas entre ellos en los diferentes momentos del tratamiento.

Si los niveles de concentración son similares en ambos tiempos para ambos fármacos, podríamos entender que existe una consistencia en la eficacia y absorción del medicamento en los pacientes. Por el contrario, si existen diferencias significativas, podrían explicar que uno de los fármacos se comporta de manera diferente o es más efectivo en ciertos pacientes o con características diferentes de éstos.

7. Conclusiones y propuestas

Después de realizar un análisis en profundidad sobre la clasificación de pacientes con tumores, se puede concluir que existen diferentes variables que apoyan y contribuyen de manera fehaciente el método de Máquinas de Soporte Vectorial (Support Vector Machine – SVM) utilizado, así como las conclusiones objetivas.

Es preciso destacar que, a pesar de todos los esfuerzos realizados, a lo largo del desarrollo del proyecto nos hemos encontrado con algunas limitaciones a la hora de proceder con la aplicación de SVM motivadas por la naturaleza no cuadrática de nuestro problema. Por ello, se han realizado diversas adaptaciones y ajustes de cara a obtener un correcto y adecuado algoritmo que se ajuste a la problemática definida según la clasificación de pacientes con tumores.

Un elemento a tener en cuenta es la presencia de pacientes cuyos valores han resultado similares en las dos características de estudio, aunque han sido tratados con procedimientos farmacológicos diferentes, por lo tanto, fueron clasificados en el grafo resultado de SVM como “mal clasificados”. Se puede afirmar que es preciso continuar con nuestra exploración ya que aún existen factores no considerados en nuestro modelo que de alguna manera pueden alterar o influenciar de manera significativa la clasificación final y por tanto los resultados.

Dado el creciente y alarmante aumento de pacientes que comienzan a padecer este tipo de tumores, es de vital importancia que la comunidad oncológica y científica continúen con las investigaciones y de alguna manera se consiga identificar y/o desarrollar uno o varios fármacos que ayuden a paliar y a mejorar tanto el estilo de vida como la supervivencia de los pacientes afectados.

Los resultados han aportado variaciones considerables en los valores sobre un mismo paciente, lo que se podría interpretar como la existencia de factores totalmente externos que pueden influir durante el desarrollo de la enfermedad, así como en la respuesta a cada tratamiento. Estas averiguaciones pueden suscitar ciertas dudas durante el proceso de medición realizado por lo que es preciso y necesario realizar un detallado seguimiento y exhaustivo análisis para cada caso.

Para finalizar, teniendo en cuenta el modelo y la metodología, han quedado sin registros o figuran de manera errónea algunas mediciones notablemente importantes, lo que supone una afección directa sobre la integridad de los datos utilizados. Es de vital importancia

realizar una recolección de datos clínicos totalmente precisa y coherente con el objetivo de conseguir una muestra de la mayor calidad que garantice un análisis preciso.

En resumen, todos nuestros hallazgos apuntan inicialmente a la naturaleza intrínseca y compleja de conseguir una óptima clasificación de pacientes con tumores, así como la necesidad de definir una amplia relación de factores para garantizar el éxito en futuras investigaciones. Es preciso continuar con la investigación y aplicación de nuevas metodologías colaborativas que, de esta manera, contribuyan a lograr novedosos avances en el tratamiento de esta enfermedad, así como mejorar el día a día de todos los afectados.

Una vez conocidos y analizados todos los hallazgos durante la aplicación de nuestro estudio, se propone implementar un modelo integrado que a su vez combine diversos enfoques y técnicas para optimizar el proceso de clasificación de pacientes con tumores, de esta manera, cualquier tratamiento será mucho más controlado y exhaustivo lo que podrá mejorar la ratio de éxito. Con ello, se pretende solucionar muchos de los problemas identificados y al mismo tiempo ordenar, analizar y aprovechar al máximo toda la información clínica disponible sobre la que tomar las mejores y más acertadas decisiones.

7.1 Propuestas

7.1.1 Modelo de Aprendizaje Profundo para características no lineales

Al encontrarnos ante un problema no cuadrático y que presenta características no lineales, se puede considerar el uso de redes neuronales profundas, como las redes neuronales convolucionales (CNN) o las redes neuronales recurrentes (RNN), cuyos modelos tienen la capacidad de analizar y aprender patrones complejos lo que ayudaría a mejorar la clasificación y por tanto la precisión de los resultados.

7.1.2 Integración de Datos Multimodales

Para no depender de datos clínicos, dada su complejidad para obtenerlos de manera rápida, sencilla y precisa, es preciso considerar la integración de diversas fuentes de información que nutran a cualquier modelo como, por ejemplo, imágenes médicas, datos de estudio del genoma, así como información sobre características hereditarias obtenidas de analíticas como ácido nucleico. Con ello, se tendrá la capacidad de obtener una representación lo más holística y detallada posible de cada paciente, ayudando de esta manera a identificar los mejores tratamientos según los marcadores tumorales identificados.

7.1.3 Aprendizaje Semi Supervisado y transferencia de aprendizaje

Debido al gran número de pacientes cuya clasificación es errónea en el grafo SVM se propone la aplicación de técnicas de aprendizaje semi-supervisado y transferencia. Dichas técnicas, comúnmente utilizadas en modelos de machine learning se basan en algoritmos que de manera iterativa generan una profunda fuente de datos etiquetados y no etiquetados que contribuyen a generar un modelo con una significativa reducción en el número de errores de clasificación.

7.1.4 Modelos de series temporales y análisis de longitudinalidad

Dadas las variaciones de datos en un mismo paciente durante el periodo de análisis, se propone utilizar la tipología de modelos de series temporales y análisis de longitudinalidad para capturar cambios y tendencias durante el tratamiento. Con toda seguridad, serán de gran ayuda para identificar patrones de respuesta y por lo tanto realizar los ajustes necesarios en el tratamiento de cada paciente de una manera totalmente dinámica y controlada.

7.1.5 Enfoque de medicina personalizada

Quizá este enfoque de investigación se convierta en el futuro como la única vía de tratamiento eficaz para este y otros muchos tipos de patologías oncológicas, ya que adaptar el tanto el modelo de tratamiento como la tipología de fármacos de manera personalizada va a suponer una auténtica revolución médica. Realizar tratamientos ad-hoc se convertirá en un modelo de éxito, todos los pacientes tienen los mismos principios en su ADN, un mismo tratamiento o fármaco puede ser totalmente eficaz o ineficaz según el sujeto al que se le administre.

Todas las propuestas anteriores nacen con el firme objetivo de ayudar a todos y cada uno de los pacientes afectados por lo que un enfoque de estudio multidisciplinar, preciso y avanzado integrando técnicas de aprendizaje, análisis multimodal y enfoques personalizados, ayudarán a mejorar el día a día de todos los pacientes, su tratamiento y por lo tanto las expectativas de supervivencia en aquellos casos cuya complejidad o avance de la enfermedad se encuentre en un estado crítico.

8. Bibliografía

- https://seom.org/publicaciones/el-cancer-en-espanyacom
- https://www.cancer.org/es/cancer/como-sobrellevar-el-cancer/tipos-de-tratamiento/quimioterapia/como-funcionan-los-medicamentos-de-quimioterapia.html
- https://www.contraelcancer.es/sites/default/files/migration/todo-sobre-cancer/tipos-cancer/cancer-ano/documentos/guia-cancer-colorrectal.pdf
- https://www.msdmanuals.com/es/hogar/c%C3%A1ncer/prevenci%C3%B3n-y-tratamiento-del-c%C3%A1ncer/principios-del-tratamiento-oncol%C3%B3gico
- https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model
- https://docta.ucm.es/rest/api/core/bitstreams/27b92c22-b95a-4992-8408-2962a474bd1e/content