Aplicación de algoritmos de inteligencia artificial para la detección de biomarcadores durante el tratamiento de pacientes con cáncer de vejiga

Andrés Palma Ponce, Pontificia Universidad Católica del Ecuador, bapalma@puce.edu.ec

***Abstract*—This study explores the application of artificial intelligence algorithms to genomic data in search of potential biomarkers for personalized bladder cancer treatments. For this study, the data was obtained from an open access repository. After the processing of the data, several algorithms were used to generate a robust model. Despite extensive analysis using advanced machine learning techniques, no significant relationship was found between altered genes and patient treatments. This underlines the complexity of underlying genetic mechanisms and the challenges in identifying effective biomarkers for oncology treatment personalization. While the initial objective was not achieved, this study emphasizes the ongoing need for research to enhance the precision and efficacy of personalized medicine in cancer treatment.**

***Keywords*—Biología computacional, Biomarcadores, Cáncer de vejiga, Inteligencia artificial, Aprendizaje de máquina**

I. Introducción

La detección de biomarcadores genéticos representa un aspecto muy importante del ámbito médico oncológico, ya que estos pueden ayudar a detectar, controlar y dar seguimiento a varios tipos de cáncer en diversos estadios. Además, en la actualidad, el auge de la inteligencia artificial ha demostrado ser de gran utilidad para el análisis de datos y realizar predicciones o clasificación de dichos datos. Para este estudio se propone la aplicación de algoritmos de inteligencia artificial para la detección de biomarcadores durante el tratamiento de pacientes con cáncer de vejiga. El objetivo principal es evaluar la efectividad de los algoritmos de inteligencia artificial al momento de detectar biomarcadores genéticos en pacientes con cáncer según el tipo de tratamiento que hayan recibido y el resultado del tratamiento.

Para lograrlo, se utilizó un conjunto de datos clínicos y moleculares recopilados de pacientes con cáncer de vejiga con distintos tratamientos. Estos datos fueron procesados y limpiados, eliminando registros nulos, estandarizando los datos y organizando la información de manera que se mantengan consistentes para el desarrollo del proyecto. Luego se aplicaron diferentes técnicas de aprendizaje automático para clasificación, de modo que se pudieron identificar patrones y características distintivas asociadas a los biomarcadores relevantes. Finalmente, se generó un modelo que muestra la efectividad de los algoritmos al momento de seleccionar genes y su relación con los diversos tratamientos y resultados de los pacientes. De encontrarse correlaciones significativas entre estos datos, podrían resultar muy útiles en futuros estudios para proponer tratamientos personalizados o encontrar relaciones entre los biomarcadores y los distintos tratamientos en cada paciente.

II. Objetivos

- Objetivo general Aplicar algoritmos de inteligencia artificial para la detección de biomarcadores durante el tratamiento de pacientes con cáncer de vejiga con el fin de mejorar la monitorización del progreso del tratamiento y contribuir a una atención médica más precisa y personalizada según el tipo de tratamiento y los biomarcadores detectados.
- Objetivos específicos
  - Analizar y procesar datos genómicos oncológicos correspondientes a genes afectados en pacientes con cáncer de vejiga para la búsqueda de biomarcadores genéticos.
  - Explorar diversos algoritmos de inteligencia artificial para identificar los biomarcadores más relevantes en el del tratamiento del cáncer de vejiga.
  - Desarrollar modelos de detección de biomarcadores utilizando los algoritmos de inteligencia artificial explorados, aplicable en el área de la biomedicina y oncología.

III. Marco teórico

A. El cáncer de vejiga

El cáncer de vejiga, también conocido como carcinoma vesical, es una afectación que se presenta cuando las células de este órgano comienzan a crecer desmesuradamente, y según los tipos de tejidos afectados y la forma en la que se presente, existen varios tipos, como puede ser el carcinoma de células escamosas o el adenocarcinoma [1]. Se considera que el origen de este cáncer es multifactorial, atribuyéndose principalmente al consumo de tabaco y exposición ocupacional, aunque también se ha encontrado que puede ser producido por factores ambientales o genéticos [2]. En cuanto a la tasa de incidencias y afectación, se ha encontrado que este tipo de cáncer es el décimo tipo de cáncer más común del mundo, siendo el sexto más común entre hombres y el decimoséptimo más común entre mujeres [3]. En un país desarrollado como Estados Unidos es la octava causa de muerte por cáncer [4], mientras que en países menos desarrollados como Egipto y Túnez, la tasa de mortalidad puede llegar a entre 5,2 y 7,8 muertes por cada 100.000 habitantes [3], por lo que el estudio de esta enfermedad aún resulta de especial relevancia en la comunidad médica y oncológica. En cuanto a los tratamientos que existen actualmente, algunos de los más comunes son la radioterapia, inmunoterapia, quimioterapia, la inmunoterapia

intravesical por medio del bacilo de Calmette-Guérin (BCG) y la cistectomía [5]. Debido a la complejidad que representa el cáncer de vejiga por los estadios, prevalencia, agravantes, sexo, edad o demografía, es importante apoyarse de varios métodos que permitan una temprana detección y un correcto tratamiento, como puede ser el uso de biomarcadores genéticos para el estudio del cáncer de vejiga [6], [7].

B. Los biomarcadores genéticos

Uno de los principales ámbitos de estudio dentro el área médica oncológica es el uso de biomarcadores genéticos. Los biomarcadores son medidas biológicas que permiten, entre otras cosas, predecir riesgos de enfermedad, mejorar la selección de tratamientos y monitorear el progreso de dicho tratamiento [8]. Los biomarcadores ayudan a determinar la presencia o progresión de una condición específica dentro del cuerpo de una persona y cómo esa condición está respondiendo al tratamiento [9]. En el caso de los biomarcadores genéticos, estos permiten detectar alteraciones a nivel de los genes, como mutaciones y pérdida de heterocigosidad en genes supresores de cáncer como el TP53 o el APC [10].

C. El uso de la inteligencia artificial en la detección de biomarcadores genéticos

En el contexto de la bioinformática y biología computacional, el uso de algoritmos de clasificación y selección ha demostrado tener un potencial en la búsqueda e identificación de biomarcadores [11]. La aplicación de algoritmos de inteligencia artificial para la detección de biomarcadores durante el tratamiento de pacientes con cáncer de vejiga se encuentra en constante evolución y se ha realizado un progreso significativo que demuestra su potencial en los últimos años [12].

Por ejemplo, en el estudio hecho por Al Abir y otros [13] se desarrolló un método de identificación de biomarcadores basado en un codificador automático invirtiendo el mecanismo de aprendizaje de los codificadores entrenados. La metodología superó todos los métodos de última generación, permitiendo detectar el tipo de cáncer con una exactitud del 99.93%, lo que confirma el potencial de los biomarcadores recientemente identificados, así como la eficacia del procedimiento de identificación de biomarcadores.

En el estudio llevado a cabo por Ma y otros [14] se utilizó el aprendizaje automático para identificar un biomarcador pronóstico de dieciséis genes para el adenocarcinoma de pulmón. El enfoque desarrollado se validó utilizando múltiples conjuntos de datos, y se encontró que los biomarcadores identificados estaban asociados con genes y eventos moleculares asociados con el cáncer.

Ge y otros en[15] analizaron las características cuantitativas llamadas “radiómicas” y biomarcadores para mostrar su potencial en detección, evaluación y seguimiento del cáncer de vejiga. Su uso, en combinación con el uso de algoritmos de aprendizaje de máquina, ha tenido un gran éxito en varios casos señalados en dicho estudio.

D. Beneficios de la detección de biomarcadores genéticos en el tratamiento del cáncer

En los últimos años ha habido un crecimiento significativo en el campo del aprendizaje automático, especialmente en los algoritmos de inteligencia artificial. Estos algoritmos tienen la capacidad de analizar grandes conjuntos de datos complejos y extraer patrones y características que pueden ser difíciles de detectar con métodos tradicionales. Aplicar estas técnicas al campo de la detección de biomarcadores en el cáncer de vejiga puede proporcionar nuevas perspectivas y enfoques para mejorar la precisión y la eficacia del diagnóstico y el tratamiento. Además, es posible extender el uso de estos algoritmos para su aplicación en otro tipo de enfermedades y tipos de cáncer, por lo que resulta un tema muy importante dentro del ámbito clínico y oncológico.

Varios estudios han demostrado que es posible identificar varios biomarcadores en pacientes con cáncer de vejiga, y que estos han permitido realizar una toma de decisiones más efectivas dentro de su tratamiento [6], [16], [17].

IV. Materiales

Los datos utilizados para el desarrollo del proyecto provienen del estudio realizado por Clinton y otros [18], donde se estudia la heterogeneidad genómica y el secuenciamiento de tumores en pacientes con cáncer de vejiga. Los datos de este estudio muestran las alteraciones que existen en los genes de los pacientes afectados por esta enfermedad que se han sometido a varios tratamientos con distintos resultados. Estos datos se encuentran alojados en cBioPortal [19]–[21]. Los *scripts* utilizados para limpieza, procesamiento, estandarización y organización de los datos fueron realizados en Python 3 [22] ya que este lenguaje cuenta con librerías especializadas como Pandas [23], que facilitan estas tareas. Para la selección de características, clasificación y creación de modelos, se utilizó la herramienta Weka [24], que posee una gama de algoritmos para la realización de estas tareas, además de una interfaz intuitiva que facilita la ejecución de varios algoritmos y creación de modelos.

V. Metodología

Para la realización del proyecto se utilizó una metodología que sigue, de manera general, los pasos básicos dentro de un proyecto de inteligencia artificial: recolección y selección de datos, preparación y preprocesamiento de datos, selección de los algoritmos para la creación y evaluación del modelo, análisis de resultados e interpretación, y la evaluación del informe [25]–[27]. Los pasos que se siguieron se explican con mayor detalle a continuación.

A. Recolección y selección de datos

El primer paso dentro de la realización del proyecto fue la selección de un conjunto de datos de cBioPortal con información de genes afectados de pacientes con cáncer de vejiga que han sido sometidos a distintos tratamientos. Estos datos deben contener información clínica que nos permita determinar aspectos como tratamientos y resultados, así como información de los genes alterados en cada muestra.

B. Preparación y preprocesamiento de datos

En esta fase realiza la exploración de los datos para su procesamiento, limpieza y estandarización. Dentro de las tareas que constan dentro de esta fase se encuentra la limpieza de datos nulos, duplicados, instancias que no pertenezcan a ninguna clase, estandarización del set de datos para obtener consistencia en cuanto los tipos de datos y la estructuración que debe tener para su posterior análisis. Para esta fase se utilizaron *scripts* creados en Python junto con la librería Pandas para facilitar el procesamiento.

Debido a la gran cantidad de genes o características (*features*) que pueden existir, un paso a seguir, posterior a la limpieza y organización de datos, es la reducción de dimensionalidad. En este paso se aplican técnicas como relación de ganancia de información, chi-cuadrado, análisis de componentes, análisis de correlación y un evaluador de envoltura (*wrapper*), específicamente el método mRMR (*Minimum Redundancy Maximum Relevance*). También es posible utilizar métodos más sencillos como es la selección basada en frecuencias [28] en ciertos casos donde las otras técnicas no tengan el desempeño esperado. Para realizar la reducción de dimensionalidad se utilizó la herramienta Weka.

C. Selección de algoritmos para la creación del modelo

La suite de Weka también provee una serie de algoritmos para la realización de modelos de clasificación. Los algoritmos evaluados para realizar la clasificación en este estudio fueron: SVM, regresión logística (LR), Naïve-Bayes (NB), clasificador multiclase (MCC), perceptrón multicapa (MLP), *random forest* (RF) y *random tree* (RT). Estos algoritmos se evaluaron bajo las configuraciones estándares provistas por Weka en un esquema de validación cruzada. Por defecto, Weka divide el conjunto de datos en 10 partes iguales. De esta manera, en cada iteración, cada grupo se transforma en el conjunto de prueba, y el resto se mantiene como conjunto de entrenamiento. Así, se asegura de que se clasifiquen todos los datos.

D. Creación y evaluación del modelo

Cada uno de los algoritmos fue evaluado para determinar su desempeño con base en métricas como fueron la exactitud (*accuracy*), la precisión, la sensibilidad, la medida F1 y el área ROC. Con base en estas métricas es posible obtener una conclusión con respecto al desempeño obtenido por cada algoritmo.

VI. Resultados

El conjunto de datos contiene 1659 muestras de 1244 pacientes. Cada muestra contiene, en primera instancia, un listado de los genes que se encuentran alterados, así como información con respecto al tipo de mutación, cambios a nivel de proteínas y frecuencia de alelos. En cuanto a la información de los pacientes, esta indica aspectos como estado de supervivencia (si está vivo o muerto al momento de tomar la muestra), tipo de muestra tomada (por ejemplo, metástasis, tumor primario, ADN circundante en el plasma) y tratamiento intravesical recibido (quimioterapia, BCG, ambos tratamientos o pacientes naïve, es decir, pacientes que no han recibido tratamiento previo). A partir de los datos seleccionados, se realizará un análisis para su procesamiento y la posterior creación del modelo.

Luego de organizar los datos, descartando muestras que no corresponden a ninguna clase, muestras que no poseen información sobre genes, y valores nulos, se obtuvo una matriz de 1248 muestras o instancias, y 511 genes que representan las características, junto con una columna adicional que representa la clase a la que pertenece dicha muestra.

Dentro del set de datos utilizado se encontraron diversos tipos de tratamientos, específicamente cuatro: quimioterapia, terapia BCG, naïve y pacientes que han recibido ambos tratamientos. Las muestras en las que no se especifican si el paciente ha recibido algún tratamiento o no, o en muestras donde no se indique el resultado del tratamiento fueron descartadas debido a que no aportaban información relevante para encontrar relación entre los genes con el tratamiento y el resultado de cada paciente. En las muestras también se muestra el resultado de supervivencia del paciente, es decir, si sobrevivió o no. A partir de estas características se definieron las clases utilizadas en el estudio: Naïve-Vivo (N_A), Naïve-Fallecido (N_D), BCG-Vivo (BCG_A), BCG-Fallecido (BCG_D), Quimioterapia-Vivo (CH_A), Quimioterapia-Fallecido (CH_D), Ambos-Vivo (BO_A), Ambos-Fallecido (BO_D). En las muestras también se encontraron 511 genes alterados, los cuales representan las características (*features*) dentro del set de datos.

La distribución de las muestras se observa en la Fig. 1.

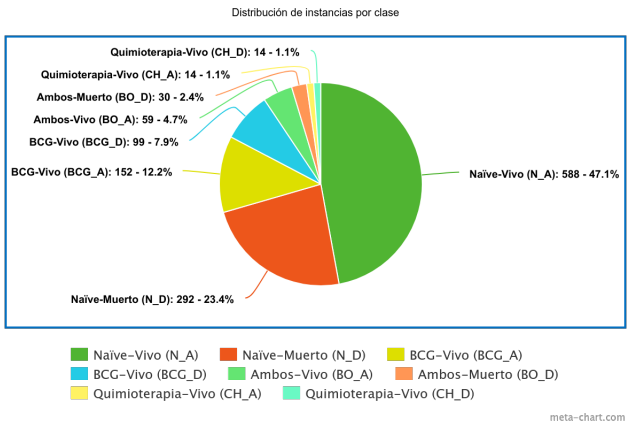

Fig. 1. Distribución de instancias en las según cada clase.

Debido a que se obtuvieron 511 características, fue necesario llevar a cabo una reducción de dimensionalidad, aplicando evaluadores como relación de ganancia de información, chi-cuadrado, análisis de componentes, análisis de correlación y con el método mRMR. Sin embargo, en ninguno de los evaluadores utilizados se encontraron coincidencias significativas en cuanto a las características más relevantes. En la Fig. 2 se muestran los resultados obtenidos de los distintos evaluadores, donde se logra observar que no existe convergencia en cuanto a las características seleccionadas como relevantes dentro del set de datos. En el caso de razón de ganancia de información, se evidencia que no se obtiene ninguna ganancia de información

por parte de las características, al igual que en el caso del evaluador chi-cuadrado. El análisis de correlación arrojó valores muy bajos para determinar una real correlación entre los genes y las clases. En el caso de PCA, la tabla muestra en la columna "*cumulative*" el porcentaje de varianza explicada por la acumulación de componentes principales y lo que se busca es llegar a un porcentaje de varianza alto, generalmente más de un 90%. En este caso, para llegar solamente a un 70% hay que ir hasta el componente 125. Como los componentes principales están conformados por una suma de pequeñas porciones de los valores de las características, es evidente que la agrupación de estos 125 componentes cubrirá una gran porción de las características, lo cual iría en contra del propósito del estudio. En el caso de mRMR, directamente no se determinó ninguna característica como relevante, pues todas obtuvieron un puntaje de 0. Finalmente, al comparar los resultados obtenidos por cada evaluador, no existe un consenso sobre las características seleccionadas por cada evaluador, por lo que no es posible determinar cuales son las características más relevantes.

Ya que no existió una correlación significativa para la selección de las principales características, siendo el desbalance de las clases una de las posibles causas, el siguiente enfoque se basó en la disminución de puntos de datos para reducir el ruido que podrían producir las clases que poseen menor número de instancias. Para ello, se seleccionó en el set de datos solo aquellas instancias pertenecientes a las clases más numerosas (N_A y N_D) para proceder ahora con una clasificación binaria. Al realizar esto se obtuvo un nuevo conjunto de datos con 880 distribuidas como se muestra en la Fig. 3, donde aún se muestra una desproporción, ya que las clases se muestran en una razón de 2 a 1.

Con este nuevo set de datos se intentó realizar la selección de características para reducir la dimensionalidad utilizando los evaluadores mencionados adnteriormente. Sin embargo, los resultados obtenidos en cuanto a puntajes y convergencia fueron muy similares a los obtenidos al realizar la reducción de dimensionalidad con todas las clases, como se muestra en la Fig. 4.

Si bien existen ciertos atributos en común entre los evaluadores de relación de ganancia de información, correlación y chi-cuadrado, se decidió complementar las características comunes entre estos evaluadores, que resulta una cantidad muy baja en comparación con todos los atributos (8 atributos en común de 511 posibles) utilizando una selección de 42 características basado en la frecuencia de alteración de cada gen, es decir, entre todas las muestras, aquellos genes que se presentan como alterados en una mayor cantidad de muestras. De estas características se seleccionaron las más frecuentes junto con aquellas que resultaron más comunes entre los anteriores evaluadores para proceder a realizar la clasificación binaria con diversos algoritmos para determinar cuáles tienen un mejor desempeño. Los genes seleccionados se muestran en la tabla I

Se procedió a realizar la clasificación utilizando una serie de algoritmos, utilizando las configuraciones por defecto provistas por la herramienta Weka. Los algoritmos evaluados muestran un desempeño bajo, con resultados de exactitud de entre 56% y 67% y área ROC entre 50% y 60.1%. En la Tabla

Relación de ganancia de información

average merit	average rank	attribute
0 +- 0	2 +- 1.61	479 PMAIP1
0 +- 0	2.4 +- 0.8	160 ASXL2
0 +- 0	3.1 +- 0.3	159 JAK1
0 +- 0	4.5 +- 0.81	161 SOX2
0 +- 0	6.2 +- 0.4	162 JAK2
0 +- 0	6.4 +- 2.15	164 IKZF1
0 +- 0	7.2 +- 0.4	158 DIS3
0 +- 0	7.5 +- 1.86	157 FLT3
0 +- 0	9.9 +- 1.81	156 MLL3
0 +- 0	11 +- 1.79	155 ATM
0 +- 0	11.2 +- 0.4	152 DNMT3B
0 +- 0	11.8 +- 0.75	153 FBXW7
0 +- 0	13 +- 0.45	154 IRS1
0 +- 0	14.4 +- 0.66	163 ARAF
0 +- 0	15.4 +- 0.66	165 EED
0 +- 0	16.4 +- 0.66	150 INSR
0 +- 0	17.6 +- 1.02	175 MDM2

Correlación

average merit	average rank	attribute
0.078 +- 0.006	1.9 +- 1.37	119 ERCC2
0.073 +- 0.005	3.4 +- 2.5	235 ERCC3
0.074 +- 0.005	3.4 +- 1.56	404 TP53BP1
0.072 +- 0.005	4 +- 2	140 LATS1
0.07 +- 0.007	5.6 +- 4.13	171 BRAF
0.063 +- 0.006	9.3 +- 5.85	253 MGA
0.062 +- 0.005	10.5 +- 5.08	385 GREM1
0.061 +- 0.005	11.6 +- 4.76	67 FOXL2
0.059 +- 0.002	12.3 +- 3.1	282 SDHA
0.059 +- 0.005	14.1 +- 6.73	65 PBRM1
0.059 +- 0.005	14.4 +- 5.43	217 AKT2
0.059 +- 0.005	14.9 +- 9.37	141 CARD11
0.059 +- 0.006	17.1 +-12.06	26 TERT
0.057 +- 0.006	19.9 +-16.06	90 ARID2
0.056 +- 0.007	21.3 +-11.35	333 GPS2

Chi-cuadrado

average merit	average rank	attribute
0 +- 0	2.5 +- 0.81	160 ASXL2
0 +- 0	3.4 +- 1.2	159 JAK1
0 +- 0	3.6 +- 4.45	479 PMAIP1
0 +- 0	4.5 +- 1.36	161 SOX2
0 +- 0	6 +- 0.77	162 JAK2
0 +- 0	6.8 +- 0.98	158 DIS3
0 +- 0	7.5 +- 1.86	157 FLT3
0 +- 0	8.1 +- 4.18	164 IKZF1
0 +- 0	10.3 +- 2	156 MLL3
0 +- 0	11.2 +- 0.4	152 DNMT3B
0 +- 0	11.4 +- 1.96	155 ATM
0 +- 0	11.6 +- 0.92	153 FBXW7
0 +- 0	12.9 +- 0.54	154 IRS1
0 +- 0	13.6 +- 2.62	163 ARAF
0 +- 0	15.7 +- 0.9	165 EED

Análisis de componentes principales (PCA)

eigenvalue	proportion	cumulative	
31.65298	0.065	0.065	0.115RRAS2+0.106FAM58A+0.105PDCD1LG2-
9.79873	0.02012	0.08512	0.177PDCD1LG2+0.175FAM58A+0.158IGF1+(
6.87587	0.01412	0.09924	-0.21PDPK1-0.148CDC42-0.139SETD8-0.1
6.26787	0.01287	0.11211	0.22 SDHC+0.17 HIST1H3A+0.154HIST1H3
5.55651	0.01141	0.12352	0.151RAC2+0.142FLCN+0.135TNFRSF14+0.
5.0947	0.01046	0.13398	0.168WWTR1-0.168SDHB-0.134GATA1+0.13
4.58584	0.00942	0.14339	-0.189ABL1+0.171SPRED1-0.143MYC+0.14
4.2621	0.00875	0.15214	-0.165TMPRSS2+0.155FUBP1-0.133AKT1+0
4.06505	0.00835	0.16049	-0.209CTLA4-0.186CDK6-0.173H3F3B-0.1
3.99594	0.00821	0.1687	0.204PNRC1+0.165DUSP4+0.153BCL6+0.14
3.67095	0.00754	0.17624	-0.171SOX2-0.145VHL-0.139FAM175A-0.1
3.61219	0.00742	0.18365	-0.212LMO1-0.124FANCA-0.123EZH2+0.11
3.54071	0.00727	0.19092	0.189RHEB+0.165GREM1+0.156HOXB13+0.1
3.41464	0.00701	0.19793	-0.162CTLA4+0.161GNAS-0.136PPM1D-0.1
3.32344	0.00682	0.20476	0.16 TMEM127+0.158EED+0.151CRKL+0.14
3.28206	0.00674	0.2115	0.157SDHD-0.157TSHR-0.139DUSP4-0.129

mRMR

number of folds (%)	attribute
0(0 %)	1 ARID1A
0(0 %)	2 RYBP
0(0 %)	3 FAT1
0(0 %)	4 FGFR4
0(0 %)	5 BLM
0(0 %)	6 CIC
0(0 %)	7 KDM6A
0(0 %)	8 PIK3CA
0(0 %)	9 NOTCH2
0(0 %)	10 DDR2
0(0 %)	11 TGFBR2
0(0 %)	12 RHOA
0(0 %)	13 SMO
0(0 %)	14 RPS6KA4
0(0 %)	15 MLL

Fig. 2. Resultados de la reducción de dimensionalidad basándose en los evaluadores descritos en el set de datos completo

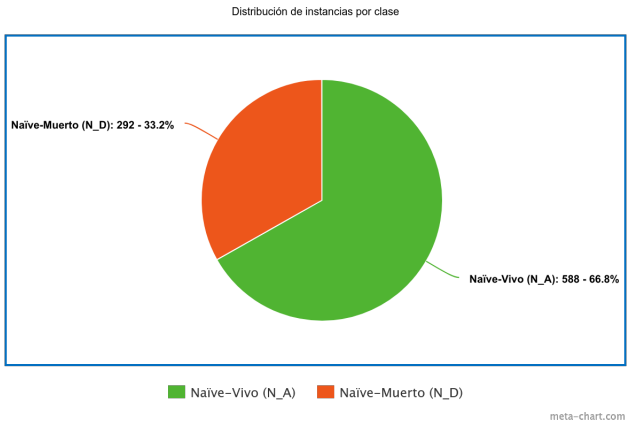

Fig. 3. Distribución de instancias en las según cada clase luego de seleccionar las clases más abundantes

.

TERT	TP53	KDM6A	ARID1A	KMT2D
FGFR3	PIK3CA	RB1	ERBB2	CREBBP
ERCC2	EP300	ATM	CDKN1A	KMT2C
FAT1	STAG2	ERBB3	ELF3	KMT2A
RBM10	TSC1	SMARCA4	SPEN	ATR
APC	PTPRT	FBXW7	NCOR1	BRCA2
ARID1B	ARID2	MGA	NSD1	NF1
ZFHX3	BRCA1	NOTCH3	ANKRD11	NOTCH4
PBRM1	PTPRD	MDC1	MITF	PRKD1
LATS1	ERCC1	AKT2	FOXL2	TP53BP1

TABLE I
LISTA DE GENES SELECCIONADOS CON BASE EN LA FRECUENCIA DE ALTERACIÓN EN LAS MUESTRAS.

II se muestran de manera detallada las métricas obtenidas por cada algoritmo evaluado.

VII. CONCLUSIONES Y RECOMENDACIONES

- En el set de datos utilizado no se encontraron relaciones significativas entre los genes alterados por cada paciente con respecto los tratamientos utilizados y resultados obtenidos. Por esta razón, no se detectaron los genes que resulten más relevantes para ser considerados como biomarcadores genéticos.
- Las herramientas utilizadas se encontraron sujetas a las limitaciones propias de cada que poseen, como una limitada gama de algoritmos, o funciones y configuraciones propias de cada herramienta. Por esta razón, un buen enfoque sería la realización del experimento aplicando algoritmos o herramientas no abordadas dentro de la investigación como el uso de redes neuronales para la clasificación, o algoritmos como BCA (Binary Correspondence Analysis), t-SNE (*t-Distributed Stochastic Neighbor Embedding*) o FAMD (*Factor Analysis of Mixed Data*) para la selección de características, ya que son métodos apropiados para datos binarios.
- Es importante notar la dificultad de acceso a datos oncológicos para realización de análisis bioinformático. Sin embargo, en trabajos futuros se podría realizar la experimentación con distintas fuentes de datos para poder contrastar los resultados. Es importante considerar otros estudios en los que posea mayor cantidad de datos o

Relación de ganancia de información

average merit	average rank	attribute
0.077 +- 0.001	2.3 +- 1.27	273 AKT2
0.073 +- 0.001	4.4 +- 2.42	351 MYOD1
0.069 +- 0.001	8.3 +- 2.49	371 FGF19
0.057 +- 0.008	12.9 +- 5.37	75 TP53BP1
0.052 +- 0.011	13.9 +- 5.49	181 ERCC3
0.045 +- 0.012	18.5 +- 7.3	234 RAD51C
0.04 +- 0.008	20.3 +- 5	66 LATS1
0.041 +- 0.008	20.6 +- 6.58	135 PRKD1
0.043 +- 0.018	22.5 +-19.48	211 MITF
0.043 +- 0.019	22.7 +-20.74	215 GREM1
0.045 +- 0.019	26.9 +-36.19	199 FOXL2
0.012 +- 0.002	27.9 +- 5.5	6 FGFR3
0.061 +- 0.02	34.1 +-73.03	402 CDKN2C
0.063 +- 0.021	34.9 +-81.77	385 CBFB
0.063 +- 0.021	35.6 +-85.85	382 IGF2

Correlación

average merit	average rank	attribute
0.121 +- 0.009	1.5 +- 0.92	75 TP53BP1
0.11 +- 0.009	2.8 +- 1.08	66 LATS1
0.11 +- 0.011	3.1 +- 1.92	6 FGFR3
0.095 +- 0.009	7.5 +- 2.73	308 STAT3
0.091 +- 0.006	8.5 +- 2.8	181 ERCC3
0.091 +- 0.008	9.5 +- 6.14	135 PRKD1
0.091 +- 0.01	10.4 +- 6.89	1 TERT
0.088 +- 0.005	10.9 +- 6.67	89 BRAF
0.089 +- 0.01	13.2 +- 7.73	11 ERCC2
0.083 +- 0.003	15.2 +- 3.84	273 AKT2
0.085 +- 0.007	15.6 +- 7.81	199 FOXL2
0.084 +- 0.009	16.4 +- 9.85	33 MGA
0.085 +- 0.012	18.2 +-13.09	229 MLL3
0.081 +- 0.007	20.4 +- 9.53	211 MITF
0.081 +- 0.009	20.9 +-11.61	215 GREM1

Chi-cuadrado

average merit	average rank	attribute
11.659 +- 1.631	1.5 +- 0.92	75 TP53BP1
9.636 +- 1.542	2.8 +- 1.08	66 LATS1
9.645 +- 1.847	3.1 +- 1.92	6 FGFR3
6.621 +- 0.853	8.1 +- 2.62	181 ERCC3
6.662 +- 1.247	8.4 +- 5.06	135 PRKD1
5.437 +- 0.436	13.2 +- 3.63	273 AKT2
4.826 +- 0.776	17.4 +- 4.67	234 RAD51C
4.936 +- 1.91	20.3 +-20.57	215 GREM1
4.064 +- 0.356	20.6 +- 4	351 MYOD1
4.884 +- 1.805	20.9 +-19.15	211 MITF
5.332 +- 1.902	22.6 +-33.58	199 FOXL2
3.154 +- 0.234	23.8 +- 3.06	371 FGF19
5.724 +- 1.947	31.5 +-69.56	89 BRAF
0.585 +- 1.754	35 +- 9.58	160 ABL1
0 +- 0	35.6 +- 6.26	163 CDKN1B

Análisis de componentes principales (PCA)

eigenvalue	proportion	cumulative	
35.79676	0.0752	0.0752	0.137RRAS2+0.137EGFL7+0.113MAPKAP1
10.71758	0.02252	0.09772	-0.145CD79A-0.144PDCD1LG2-0.143FAM
7.6914	0.01616	0.11388	0.192CDK4-0.189SDHC-0.163HIST1H3A-
6.8466	0.01438	0.12826	-0.214CDK4-0.179NTHL1+0.171GREM1-0
5.85203	0.01229	0.14056	0.203SDHB+0.178BABAM1+0.143VTCN1+0
5.49996	0.01155	0.15211	-0.163MST1-0.154ABL1+0.149SPRED1-0
5.15056	0.01082	0.16293	0.183FUBP1-0.155TMPRSS2-0.14AKT1+0
4.7774	0.01004	0.17297	-0.217CTLA4-0.217CDK6-0.211H3F3B-0
4.33412	0.00911	0.18207	-0.3LMO1-0.152CRKL-0.138TGFBR2-0.1
4.29263	0.00902	0.19109	-0.172RHEB-0.136HOXB13-0.134RAD51I
4.07129	0.00855	0.19964	-0.201RHEB-0.152GREM1+0.139NPM1+0.
4.02046	0.00845	0.20809	-0.155VEGFA+0.153SMYD3-0.147IRF4+0
3.85994	0.00811	0.2162	0.162HOXB13+0.152MRE11A+0.141NBN+0
3.75273	0.00788	0.22408	0.133TMEM127+0.129SOCS1+0.122CCND2
3.67214	0.00771	0.2318	-0.151HLA-A+0.139CRKL-0.136CSDE1+0

mRMR

number of folds (%)	attribute
0(0 %)	1 TERT
0(0 %)	2 TP53
0(0 %)	3 KDM6A
0(0 %)	4 ARID1A
0(0 %)	5 KMT2D
0(0 %)	6 FGFR3
0(0 %)	7 PIK3CA
0(0 %)	8 RB1
0(0 %)	9 ERBB2
0(0 %)	10 CREBBP
0(0 %)	11 ERCC2
0(0 %)	12 EP300
0(0 %)	13 ATM
0(0 %)	14 CDKN1A
0(0 %)	15 KMT2C

Fig. 4. Resultados de la reducción de dimensionalidad basándose en los evaluadores descritos en el conjunto de datos con solo con dos clases

Algoritmo	Exactitud (*Accuracy*)	Precisión	Sensibilidad (*Recall*)	F1-Measure	Área ROC
SVM	66.81%	66.8%	100%	80.1%	50%
LR	64.43%	67.9%	84.2%	76.9%	54.8%
MLP	63.63%	71.1%	76.9%	73.9%	59.4%
NB	56.36%	68.0%	65.5%	66.7%	54.3%
MCC	64.43%	67.9%	88.6%	76.9%	54.8%
RF	67.84%	70.9%	88.1%	78.5%	60.1%
RT	59.54%	68.4%	73.3%	70.8%	52.4%

TABLE II
RESULTADOS EN LA CLASIFICACIÓN BINARIA CON DISTINTOS ALGORITMOS

donde se tomen en cuenta otro tipo de factores como evolución del tratamiento en diferentes estadios.
- Es importante profundizar en los fundamentos biológicos e informáticos para comprender la relevancia de los datos y la selección de los algoritmos más apropiados para realizar las pruebas.

VIII. TRABAJO FUTURO

Debido a ciertas limitaciones al momento de realizar el estudio, es importante destacar posibles mejoras o extensiones que pueden realizarse en futuros trabajo a partir de este

- La adquisición de datos oncológicos para análisis bioinformático puede resultar complicada, limitando la capacidad y alcance del estudio, razón por lo cual, en caso de tener la posibilidad de obtener conjuntos de datos similares, replicar y complementar el estudio con el fin de contrastar los resultados obtenidos.
- Para este estudio, para mayor simplicidad al momento de analizar diversos algoritmos, se utilizaron herramientas como Weka, que en ciertas situaciones puede restringir la libertad sobre los algoritmos que posee. Por esta razón, es conveniente replicar el estudio utilizando otros algoritmos u otras herramientas que provean mayor flexibilidad y control sobre los algoritmos con los que se puede realizar el estudio, como scikit-learn y Pytorch en Python. Un buen enfoque en variar entre las configuraciones de cada algoritmo o utilizar técnicas más avanzadas como redes neuronales.
- Existen algunas razones por las cuales un set de datos podría no resultar apropiado para realizar un tipo de estudio, como puede ser un desbalance entre las clases, o atributos que no aporten valor al estudio. Por esta razón es importante analizar el efecto del desbalance de datos en este tipo de estudios, así como la relevancia de los atributos a considerar al momento de elegir un dataset.

BIBLIOGRAFÍA

[1] [Online]. Available: https://www.cancer.gov/espanol/tipos/vejiga.

[2] J. L. Brito, A. Á. Llergo, L. M. C. Pérez, and R. M. M. Jiménez, "Revisión sistemática sobre el cáncer de vejiga y exposición ocupacional," *Medicina Y Seguridad Del Trabajo*, vol. 66, no. 259, pp. 81–99, Apr. 2020. DOI: 10.4321/s0465-546x2020000200003. [Online]. Available: https://doi.org/10.4321/s0465-546x2020000200003.

[3] W. International, *Bladder cancer statistics — world cancer research fund international*, en-US, Apr. 2022. [Online]. Available: https://www.wcrf.org/cancer-trends/bladder-cancer-statistics/.

[4] Mar. 2023. [Online]. Available: https://www.cancer.net/cancer-types/bladder-cancer/statistics.

[5] [Online]. Available: https://www.cancer.gov/espanol/tipos/vejiga/tratamiento.

[6] I. Proctor, K. Stoeber, and G. Williams, "Biomarkers in bladder cancer," *Histopathology*, vol. 57, no. 1, pp. 1–13, Jul. 2010. DOI: 10.1111/j.1365-2559.2010.03592.x. [Online]. Available: https://doi.org/10.1111/j.1365-2559.2010.03592.x.

[7] F. Ye, L. Wang, M. Castillo-Martin, *et al.*, "Biomarkers for bladder cancer management: Present and future," *American Journal of Clinical and Experimental Urology*, vol. 2, no. 1, pp. 1–14, 2014. DOI: 10.1016/j.cbi.2021.12.054.

[8] J. Simon, "Genomic biomarkers in predictive medicine. an interim analysis," *EMBO Molecular Medicine*, vol. 3, no. 8, pp. 429–435, Jul. 2011. DOI: 10.1002/emmm.201100153. [Online]. Available: https://doi.org/10.1002/emmm.201100153.

[9] CancerCare, *Role of biomarkers — cancer*. [Online]. Available: https://www.cancercare.org/publications/413-understanding_the_role_of_biomarkers_in_treating_cancer.

[10] J. M. Carethers and B. Jung, "Genetics and genetic biomarkers in sporadic colorectal cancer," *Gastroenterology*, vol. 149, no. 5, 1177–1190.e3, Oct. 2015. DOI: 10.1053/j.gastro.2015.06.047. [Online]. Available: https://doi.org/10.1053/j.gastro.2015.06.047.

[11] X. Zhang, I. Jonassen, and A. Goksøyr, "Machine learning approaches for biomarker discovery using gene expression data," in *Exon Publications eBooks*. Mar. 2021, pp. 53–64. DOI: 10.36255/exonpublications.bioinformatics.2021.ch4. [Online]. Available: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4.

[12] Z. Jagga and D. Gupta, "Machine learning for biomarker identification in cancer research – developments toward its clinical application," *Personalized Medicine*, vol. 12, no. 4, pp. 371–387, Aug. 2015. DOI: 10.2217/pme.15.5. [Online]. Available: https://doi.org/10.2217/pme.15.5.

[13] F. Al Abir, S. Shovan, M. A. M. Hasan, A. Sayeed, and J. Shin, "Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination," *Molecular Omics*, vol. 18, no. 7, pp. 652–661, 2022.

[14] B. Ma, Y. Geng, F. Meng, Y. Ge, and F. Song, "Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method," *Journal of Cancer*, vol. 11, no. 5, pp. 1288–1298, Jan. 2020. DOI: 10.7150/jca.34585. [Online]. Available: https://doi.org/10.7150/jca.34585.

[15] L. Ge, Y. Chen, C. Yan, *et al.*, "Study progress of radiomics with machine learning for precision medicine in bladder cancer management," *Frontiers in Oncology*, vol. 9, Nov. 2019. DOI: 10.3389/fonc.2019.01296. [Online]. Available: https://doi.org/10.3389/fonc.2019.01296.

[16] J.-H. Ahn, C.-K. Kang, E.-M. Kim, A.-R. Kim, and A. Kim, "Proteomics for early detection of non-muscle-invasive bladder cancer: Clinically useful urine protein biomarkers," *Life*, vol. 12, no. 3, p. 395, 2022.

[17] R. Chou, J. L. Gore, D. I. Buckley, *et al.*, "Urinary biomarkers for diagnosis of bladder cancer," *Annals of Internal Medicine*, vol. 163, no. 12, pp. 922–931, Dec. 2015. DOI: 10.7326/m15-0997. [Online]. Available: https://doi.org/10.7326/m15-0997.

[18] T. Clinton, Z. Chen, H. Wise, *et al.*, "Genomic heterogeneity as a barrier to precision oncology in urothelial cancer," *Cell Reports*, vol. 41, no. 12, p. 111859, Dec. 2022. DOI: 10.1016/j.celrep.2022.111859. [Online]. Available: https://doi.org/10.1016/j.celrep.2022.111859.

[19] E. Cerami, J. Gao, U. Doğrusöz, *et al.*, "The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, pp. 401–404, May 2012. DOI: 10.1158/2159-8290.cd-12-0095. [Online]. Available: https://doi.org/10.1158/2159-8290.cd-12-0095.

[20] J. Gao, B. A. Aksoy, U. Doğrusöz, *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal," *Science Signaling*, vol. 6, no. 269, Apr. 2013. DOI: 10.1126/scisignal.2004088. [Online]. Available: https://doi.org/10.1126/scisignal.2004088.

[21] I. De Bruijn, R. Kundra, B. Mastrogiacomo, *et al.*, "Analysis and visualization of longitudinal genomic and clinical data from the aacr project genie biopharma collaborative in cbioportal," *Cancer Research*, vol. 83, no. 23, pp. 3861–3867, Sep. 2023. DOI: 10.1158/0008-5472.can-23-0816. [Online]. Available: https://doi.org/10.1158/0008-5472.can-23-0816.

[22] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[23] T. pandas development team, *Pandas-dev/pandas: Pandas*, version latest, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: https://doi.org/10.5281/zenodo.3509134.

[24] F. Eibe, M. A. Hall, and I. H. Witten, "The weka workbench. online appendix for data mining: Practical machine learning tools and techniques," in *Morgan Kaufmann*, Morgan Kaufmann Publishers San Francisco, California, 2016.

[25] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20 991–21 002, 2020. DOI: 10.1109/ACCESS.2019.2963053.

[26] W. Wu, Y.-J. Li, A. Feng, *et al.*, "Data mining in clinical big data: The frequently used databases, steps, and methodological models," *Military Medical Research*, vol. 8, no. 1, Aug. 2021. DOI: 10.1186/s40779-021-00338-z. [Online]. Available: https://doi.org/10.1186/s40779-021-00338-z.

[27] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, en, Second Edition. "O'Reilly Media, Inc.", Oct. 2022.

[28] C. D. Manning, *An introduction to information retrieval*. Cambridge university press, 2009.