Identificación de patrones químico-biológicos en la detección de Cáncer de Próstata utilizando técnicas de machine learning

RESUMEN: El Cáncer de Próstata es considerada la segunda causa en mortalidad entre varones, existen dos pruebas de inicio que permite detectarlo entre ellas se encuentras la Prueba de Antígeno Prostático (PSA) y el tacto rectal, estos son considerados como métodos invasivos; los cuales una vez realizados se requiere otros estudios que confirmen de si padece o no la enfermedad, tales como; resonancias magnéticas y biopsias. Por tal motivo al usar inteligencia artificial y obtener un modelo predictivo coadyuvaría en la detección del cáncer de próstata.

En este proyecto se utilizaron herramientas de inteligencia artificial por medio de subrutinas en Python como Scikit-learn, Pandas, Seaborn y Matplotlib para el análisis de una base de datos de 985 pacientes con Cáncer de Próstata, con el fin de determinar las correlaciones tanto biológicas y químicas para su detección. Las variables que se usaron durante la investigación fueron edad, raza, el puntaje de Gleason, mutaciones, tiempo de sobrevivencia, el tipo de muestra de la biopsia y estadio, en donde se aplicó el método de correlaciones lineales. Enseguida se obtuvieron modelos predictivos basados en regresión logística, bosques aleatorios y potenciación extrema de gradiente. En los resultados se observó que entre mayor era la etapa del cáncer disminuía el tiempo para sobrevivir y de igual manera cuando el PSA se encontraba mayor a 4 ng/mL de sangre el paciente presentaba cáncer de próstata además que tenía una menor probabilidad de sobrevivir.

Además, se obtuvieron métricas estadísticas como precisión, sensibilidad, especificidad y área bajo la curva (AUC), resultando el mejor modelo predictivo basado en potenciación extrema de gradiente.

PALABRAS CLAVE: Aprendizaje automático, cáncer de próstata, inteligencia artificial, PSA.

Colaboración
Fernanda Ivette Inurreta Cortes; Marco Gallo, Tecnológico Nacional de México / Instituto Tecnológico de Ciudad Juárez

Fecha de recepción: 04 de abril de 2025
Fecha de aceptación: 29 de mayo de 2025

ABSTRACT: Prostate cancer is considered the second leading cause of death among men. There are two initial tests that allow for its detection: the Prostate Cancer Antigen (PSA) test and the digital rectal examination. These are considered invasive methods; once performed, further studies are required to confirm whether or not the patient has the disease, such as MRIs and biopsies. Therefore, using artificial intelligence and obtaining a predictive model would aid in the detection of prostate cancer.

This project used artificial intelligence tools through Python subroutines such as Scikit-learn, Pandas, Seaborn, and Matplotlib to analyze a database of 985 patients with prostate cancer, in order to determine both biological and chemical correlations for its detection. The variables used in the study were age, race, Gleason score, mutations, survival time, biopsy sample type, and stage, where the linear correlation method was applied. Predictive models based on logistic regression, random forests, and extreme gradient boosting were then obtained. The results showed that the higher the stage of the cancer, the shorter the survival time, and similarly, when the PSA was greater than 4 ng/mL of blood, the patient had prostate cancer and had a lower probability of survival.

In addition, statistical metrics such as accuracy, sensitivity, specificity and area under the curve (AUC) were obtained, resulting in the best predictive model based on extreme gradient boosting XGBoost.

KEYWORDS: Artificial intelligence, cancer of prostate, machine learning, PSA.

INTRODUCCIÓN

La próstata es una glándula que compone el aparato reproductor masculino, donde se sitúa debajo de la vejiga urinaria, tiene como función la fabricación de la proteína llamada antígeno especifico de la próstata (PSA), lo que permite sustancias que proporcionan nutrientes y protección para los espermatozoides al momento de la eyaculación, que ayuda para el surgimiento de la fertilización del ovulo [1].

Uno de los padecimientos relacionados que puede afectar esta parte de los varones es el cáncer de próstata que es considerada de las principales causas de morbilidad y mortalidad en México, esta enfermedad aparece cuando se forman células malignas (cancerosas), los síntomas se pueden presenciar en una etapa avanzada comúnmente, debido que es un cáncer que se propagan lentamente e inclusive esparcirse a otros órganos o huesos. Dentro de los signos que se presenta es a menudo evacuar la vejiga principalmente durante la noche, dificultad para iniciar o mantener el flujo urinario, molestias o escozor al orinar, problemas para eyacular y presencia de sangre en la orina o semen [2]. Para clasificar los estadios del cáncer de próstata se divide de acuerdo con el sistema de TNM de estadificación de cáncer de próstata, el cual es usado mundialmente y sus siglas son: T (Tumor primario), N (Nódulos Linfáticos Regionales) y M (Metástasis a distancia). Entre los métodos utilizados para su detección que existen son invasivos para los pacientes como es el medir el puntaje de Gleason por medio de toma de muestra de sangre o por biopsia, inclusive el tacto rectal [3].

Dentro de los campos de la medicina se está empleando la inteligencia artificial (IA) debido a la capacidad de los ordenadores para analizar grandes volúmenes de información en tiempos muy cortos, desarrollando modelos predictivos de alta precision [4], además a diferencia de las personas no requiere descanso y es posible reducir errores.

En la IA se encuentra una rama llamada aprendizaje automático, la cual permite que por medio de algoritmos, en donde los datos o variables de entrada son alimentados, obteniéndose un modelo empírico o función análoga que mapee los datos de entrada con la variable objetivo o de salida, una especie de regresión no-lineal, obteniéndose predicciones con cierto grado de confiabilidad en conjunto con análisis estadísticos [5].

En el ámbito del aprendizaje automático, se distinguen dos enfoques principales. El primero es el aprendizaje supervisado, donde un algoritmo procesa un conjunto de datos previamente etiquetados y se entrena para asignar la etiqueta adecuada a nuevos datos. Ejemplos de este tipo de algoritmos incluye la clasificación Naive Bayes, la regresión lineal y logística, las máquinas de vectores de soporte (SVM) y los árboles de decisión. Por otro lado, el aprendizaje no supervisado se basa en la identificación de patrones dentro de un conjunto de datos, usualmente sin asignar etiquetas a las muestras [6]. De manera que la hipótesis en este trabajo consiste en la identificación de patrones para la detección de Cáncer de Próstata por medio de modelos descriptivos de inteligencia artificial como regresión logística, bosque aleatorio y potenciación extrema de gradiente con una precisión adecuada.

Además otra forma en que la IA puede ayudar en la detección de cáncer de próstata (CP) es utilizando el valor de PSA como entrada para un algoritmo y determinando el riesgo de tener una biopsia de próstata positiva. Dichos algoritmos a menudo incluyen otros datos de entrada además del nivel de PSA, como, por ejemplo, la edad del paciente, el recuento de glóbulos blancos en el análisis de orina, el volumen de la próstata (estimado) y el estado del tacto rectal (DRE). Los resultados muestran que simplemente incluyendo esta información adicional del paciente y entrenando una red neuronal para obtenerse un modelo experto en la clasificación de casos de próstata, se puede lograr una mayor sensibilidad en comparación con una prueba de PSA regular [7].

También la IA se puede utilizar usando la segmentación de glándula prostática en imágenes de ultrasonido transrectal (TRUS), al proporcionar el volumen de la próstata de una manera más rápida y objetiva a diferencia de las mediciones manuales. Esto puede resultar de gran beneficio para los cálculos de densidad de PSA [8]. Además, la localización de la lesión o de la región de interés (ROI) aportaría más elementos para respaldar las biopsias. Los métodos basados en IA se han investigado en el pasado para ecografías únicas y multiparamétricas, en donde se ha demostrado que pueden detectar el cáncer de próstata y señalar los casos clínicamente significativos [9]. Como es el caso del modelo de deformación creado por Onofrey en la Universidad de Yale, al garantizar que los urólogos tomen sus muestras de biopsia de las secciones correctas de la glándula prostática, en lugar de al azar [10].

Por lo tanto, el objetivo principal de esta investigación es identificar las principales variables químico-biológicas en una base de datos con la finalidad de obtener sus correlaciones y el empleo de modelos predictivos que coadyuven en la detección de Cáncer de Próstata utilizando librerías de Python.

MATERIALES Y MÉTODOS

Para el desarrollo del proyecto se utilizaron los siguientes materiales:
Equipo de cómputo: Se requirió para el análisis y procesamiento de los resultados, el cual cuenta con las siguientes características:

Tabla 1: Materiales a utilizar en el Proyecto de Investigación.

ELEMENTO	CARACTERISTICA
Procesador	Intel(R) Core(TM) i3-N305 1.80 GHz
Memoria RAM	8.00 GB
Sistemas operativos	Windows

Fuente: Elaboración propia.

1. Paquete computacional Microsoft Office Word: Para la elaboración de texto
2. Paquete computacional Microsoft Office Excel: Para la elaboración de base de datos.
3. Google Colab: Es una plataforma que permite ejecutar códigos de Python por medio de notebooks de Jupyter (Google. (n.f.) [11]
4. cBioPortal: Es una base de datos de libre acceso a información de pacientes con diferentes tipos de cáncer (cBioPortal. (2024). Prostate cancer MSK study [Conjunto de datos]) [12-14].
5. Subrutinas de Python: Scikit-learn, Pandas, Seaborn y Matplotlib. [15-18]

Selección de Base de Datos

Durante la búsqueda inicial se encontró una base de datos del Reino Unido, sin embargo, por medio de correo electrónico se solicitó acceso, y nunca se obtuvo una respuesta. Después se encontró una base de datos por parte de NCI (National Cancer Institute), la cual solo contenía la información de 43 pacientes con cáncer de próstata, por lo que la cantidad de datos era demasiado pequeña para el estudio, y se optó por no utilizarla [19-20].

Al final se seleccionó la base de datos de cBioPortal [12-14] esta plataforma asistida por el Memorial Sloan Kettering Cancer Center, NCI, Dana-Farber Cancer Institute y Harvard, entre otras universidades, contiene información de diferentes tipos de cáncer, es de libre acceso y en ella se encontró información acerca de 2,257 pacientes con padecimiento de cáncer de próstata.

Análisis de base de datos

Una vez seleccionada la base de datos, se analizaron los datos únicamente de aquellos que contenía toda la información como edad, raza, etnia, alteraciones genéticas, mutaciones, tipo de cáncer, medida de Gleason, al igual que el tiempo que tenia de vida y también si se encontraba vivos o difuntos, al final se utilizó el expediente de 985 personas.

En seguida se utilizó Google Colab para el análisis de los datos obtenidos, primero se convirtieron las variables categóricas en numéricas, y después de determino la matriz de correlación entre variables utilizando el coeficiente de Pearson por medio de la covarianza y desviación estándar, por medio de la librería de python Pandas. Después se determianron los siguientes modelos de clasificación binaria para la variable objetivo (estado de sobrevivencia global) utilizando subrutinas de python Scikit-learn. Los modelos predictivos obtenidos en este trabajo son bosque aleatorio, regresión logística y potenciación extrema de gradiente. Los resultados se graficaron utilizando las librerías de python Seaborn y Matplotlib.

RESULTADOS

Los resultados obtenidos en el análisis de la supervivencia global por estadio de cáncer muestran una clara relación inversa entre el estadio de la enfermedad y el tiempo de supervivencia de los pacientes. Como se observa en el diagrama de cajas en la Figura 1, los pacientes diagnosticados en estadio 0 y 1 presentan una mediana de supervivencia considerablemente mayor en comparación con aquellos diagnosticados en estadio 2.

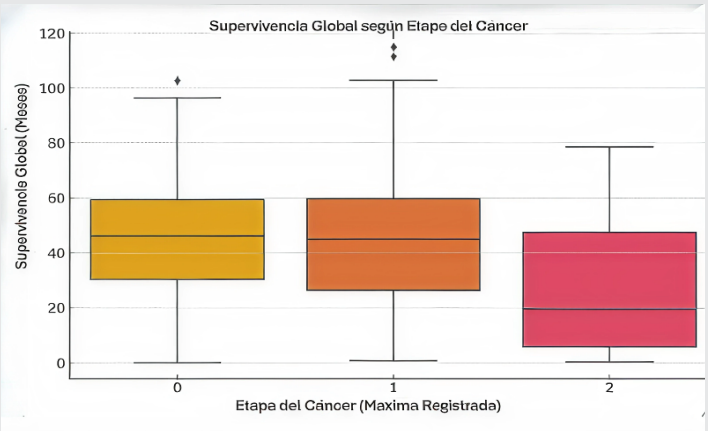

Figura 1, Distribución de la supervivencia general en meses de pacientes según la etapa más alta de cáncer registrada (0, 1 y 2).

El hecho de que la mediana de supervivencia para los estadios 0 y 1 sea similar sugiere que, en las primeras fases de la enfermedad, los tratamientos disponibles pueden ser igualmente efectivos en prolongar la vida del paciente. Sin embargo, en estadio 2, la mediana de supervivencia disminuye drásticamente, lo que sugiere una progresión más agresiva de la enfermedad o una menor eficacia terapéutica en esta fase.

Además, la variabilidad en la supervivencia dentro de cada grupo es un factor relevante. En los estadios 0 y 1, la distribución de los datos muestra valores atípicos por encima del tercer cuartil, lo que indica que ciertos pacientes logran una supervivencia significativamente mayor que el promedio. Esto podría estar relacionado con factores como la respuesta individual al tratamiento, la presencia de comorbilidades, o diferencias en la detección temprana. En contraste, el estadio 2 muestra una menor cantidad de valores atípicos, lo que sugiere una menor variabilidad y una mayor consistencia en los tiempos de supervivencia reducidos.

El análisis de la relación entre la puntuación de Gleason y la supervivencia global muestra una correlación inversa entre la agresividad del tumor y el tiempo de supervivencia de los pacientes. Como se observa en el diagrama de cajas en la Figura 2, los pacientes con puntuaciones de Gleason más bajas (6-7) presentan una mediana de supervivencia significativamente mayor en comparación con aquellos con puntuaciones más altas (9-10).

Los resultados sugieren que en los grupos con puntuación de Gleason 6 y 7, la mediana de supervivencia es relativamente alta, con una amplia distribución de datos que incluye valores atípicos en el rango superior. Esto indica que algunos pacientes logran sobrevivencias prolongadas, posiblemente debido a una detección temprana y un tratamiento efectivo. En contraste, en los pacientes con Gleason 9 y 10, la mediana de supervivencia se reduce drásticamente y la variabilidad en los tiempos de supervivencia es menor, sugiriendo un pronóstico más desfavorable y una progresión más rápida de la enfermedad.

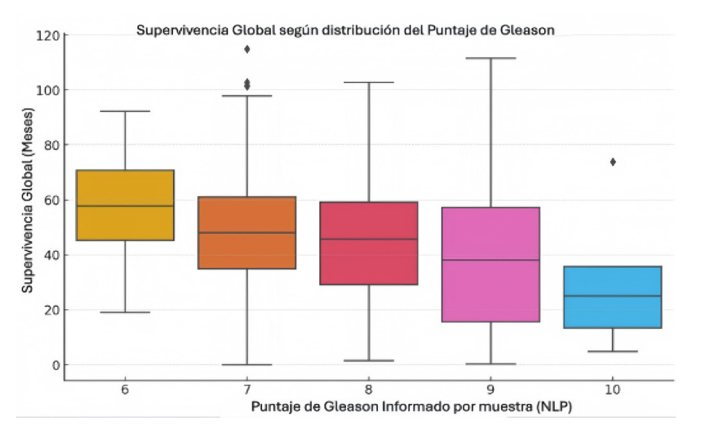

Figura 2. Relación entre la puntuación de Gleason (de 6 a 10) y la supervivencia general en meses.

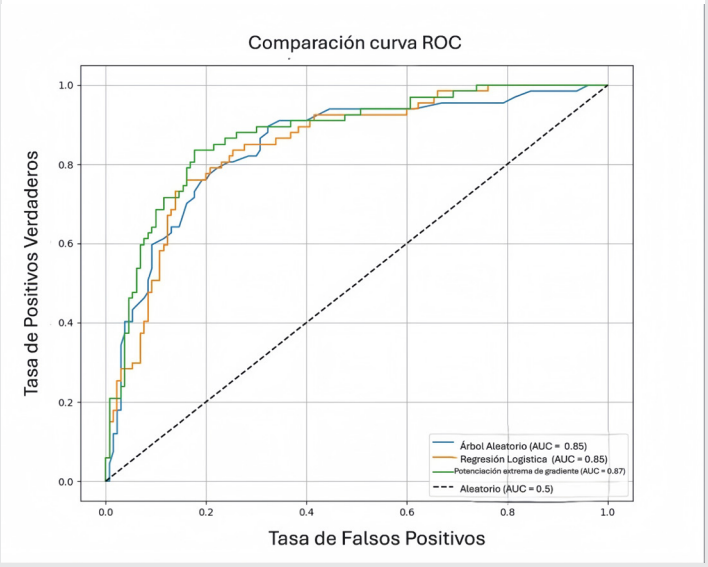

Figura 3. Curva de ROC de los modelos predictivos al detectar pacientes positivos con Cáncer de Próstata.

Para evaluar el desempeño de los modelos predictivos se analizaron cuatro métricas fundamentales que fueron la precisión, sensibilidad, especificidad y el área de bajo de la curva ROC (AUC), estas métricas nos permiten valorar la capacidad general del modelo para clasificar correctamente, como su desempeño de verdaderos positivos y verdaderos negativos. En la Figura 3, se observa que el modelo predictivo con mayor precisión para la detección de pacientes con cáncer de próstata es el de potenciación extrema de gradiente, debido que su valor de AUC=0.874, a diferencia de los otros modelos que presenta un valor de AUC= 0.85.

En la Tabla 2 se observa que el modelo de Potenciación de Gradiente Extrema muestra el mejor rendimiento general, destacándose en precisión, sensibilidad y AUC en comparación con los otros dos modelos, clasificando de forma correcta los casos tanto positivos como negativos.

Aun cuando el modelo de Bosque Aleatorio presenta la mayor especificidad de entre los tres modelos, esto viene a costa de una menor sensibilidad, lo que podría no ser ideal en contextos donde los falsos negativos al momento de detectar el cáncer de próstata. Por otro lado, la Regresión Logística, aunque tiene un rendimiento competitivo, se queda ligeramente detrás de potenciación extrema de gradiente en todas las métricas clave.

Tabla 2. Métricas estadísticas de los modelos predictivos para detección Cáncer de Próstata en los pacientes.

Tipo de Modelo	Precisión	Sensibilidad	Especificidad	AUC
Bosque Aleatorio	0.797	0.582	0.908	0.847
Regresión logística	0.787	0.597	0.885	0.848
XGBoost	0.827	0.687	0.900	0.874

Fuente: Elaboración propia.

CONCLUSIONES

Los pacientes diagnosticados en estadios iniciales de cáncer y aquellos con puntuaciones de Gleason más bajas presentan tiempos de supervivencia significativamente mayores. Además, la presencia de valores atípicos en estos grupos podría indicar que algunos individuos pueden beneficiarse de estrategias terapéuticas personalizadas o de factores individuales favorables. En contraste, los pacientes en estadios más avanzados y con puntuaciones de Gleason elevadas muestran una mediana de supervivencia reducida y menor variabilidad en los resultados, lo que indica un pronóstico más reservado.

La curva ROC evidencian un desempeño robusto de los tres modelos evaluados: Árbol aleatorio, Regre-

sión Logística y Potenciación extrema de gradiente. Tanto el modelo de Árbol Aleatorio como el de Regresión Logística alcanzaron un AUC de 0.85, mientras que Potenciación extrema de gradiente superó ligeramente a ambos con un AUC de 0.87. Estos valores indican una alta capacidad discriminativa para todos los modelos, siendo potenciación extrema de gradiente el más eficaz en términos de la relación entre la tasa de verdaderos positivos y la de falsos positivos.

Los resultados en este trabajo señalan la necesidad de continuar investigando estrategias de detección temprana y el desarrollo de tratamientos más efectivos para mejorar la supervivencia en pacientes con tumores agresivos. En este sentido, los algoritmos de inteligencia artificial están emergiendo como herramientas prometedoras para la detección temprana del cáncer, permitiendo el análisis de grandes volúmenes de datos clínicos y radiológicos para identificar patrones asociados con el desarrollo tumoral. La implementación de estos métodos podría mejorar la precisión del diagnóstico, optimizar la selección de tratamientos y en última instancia, aumentar la supervivencia de los pacientes. Permitiendo emplearse en otras líneas de investigación, desarrollando otros modelos predictivos, considerando además factores genéticos, ambientales y estilos de vida que explique valores atípicos en los tiempos de supervivencia.

BIBLIOGRAFÍA

[1] Taguchi, Y. (2009). La próstata: todo lo que necesita saber sobre la glándula masculina (2ª ed.). Amat Editorial.

[2] Instituto de Salud para el Bienestar. (s.f.). Día Mundial del Cáncer de Próstata. Recuperado el 11 de junio de 2022, de https://www.gob.mx/insabi/articulos/dia-mundial-del-cancer-de-prostata-11-de-junio?idiom=es.

[3] National Cancer Institute. (s.f.). Tratamiento del cáncer de próstata (PDQ®). Recuperado el 11 de abril de 2025, de https://www.cancer.gov/espanol/tipos/prostata/pro/tratamiento-prostata-pdq.

[4] Rouhiainen, L. (2018). Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro. Alienta Editorial.

[5] Norman, A. T. (2019). Aprendizaje automático en acción. Litres.

[6] Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020, noviembre). Machine learning methods in drug discovery. Molecules, 25(22), 5277. https://doi.org/10.3390/molecules25225277.

[7] Stephan, C., Cammann, H., Semjonow, A., Diamandis, E. P., Wymenga, L. F., Lein, M., Sinha, P., Loening, S. A., & Jung, K. (2002, agosto). Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies. Clinical Chemistry, 48(8), 1279–1287. https://doi.org/10.1093/clinchem/48.8.1279.

[8] Kachouie, N. N., & Fieguth, P. (2007, agosto). A medical texture local binary pattern for TRUS prostate segmentation. Conference Proceedings, 4225, 5605–5608. https://doi.org/10.1109/iembs.2007.4353617.

[9] Wildeboer, R. R., Mannaerts, C. K., Van Sloun, R. J. G., Budäus, L., Tilki, D., Wijkstra, H., Salomon, G., & Mischi, M. (2019, octubre). Automated multiparametric localization of prostate cancer based on B-mode, shear-wave elastography, and contrast-enhanced ultrasound radiomics. European Radiology, 30(2), 806–815. https://doi.org/10.1007/s00330-019-06436-w.

[10] Dee, J. E. (2022, mayo 10). Using AI and machine learning for a more accurate prostate biopsy. Yale School of Medicine. Recuperado el 10 de mayo de 2022, de https://medicine.yale.edu/news-article/using-ai-and-machine-learning-for-a-more-accurate-prostate-biopsy/

[11] Google Research. (s.f.). Google Colaboratory. Recuperado de https://colab.research.google.com/.

[12] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012, mayo). The CBIO Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discovery, 2(5), 401–404. https://doi.org/10.1158/2159-8290.cd-12-0095.

[13] De Bruijn, I., Kundra, R., Mastrogiacomo, B., Tran, T. N., Sikina, L., Mazor, T., Li, X., Ochoa, A., Zhao, G., Lai, B., Abeshouse, A., Baiceanu, D., Ciftci, E., Dogrusoz, U., Dufilie, A., Erkoc, Z., Lara, E. G., Fu, Z., Gross, B., ... Schultz, N. (2023, septiembre). Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. Cancer Research, 83(23), 3861–3867. https://doi.org/10.1158/0008-5472.can-23-0816.

[14] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander,

C., & Schultz, N. (2013, abril). Integrative analysis of complex cancer genomics and clinical profiles using the CBioPortal. Science Signaling, 6(269). https://doi.org/10.1126/scisignal.2004088.

[15] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55.

[16] McKinney, W. (2010, junio). Data structures for statistical computing in Python. Proceedings of the Python in Science Conferences, 56–61. https://doi.org/10.25080/majora-92bf1922-00a.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2012, octubre). Scikit-learn: Machine learning in Python. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1201.0490.

[18] Waskom, M. (2021, abril). seaborn: Statistical data visualization. The Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021.

[19] National Cancer Institute. (s.f.). Clinical Data Commons. Recuperado el 17 de noviembre de 2024, de https://clinical.datacommons.cancer.gov/#/explore.

[20] UK Biobank. (s.f.). UK Biobank. Recuperado el 17 de noviembre de 2024, de https://www.ukbiobank.ac.uk/.