

Mejora en la Detección Temprana del Cáncer Mediante Análisis de Sangre Multianalito y Modelos de Aprendizaje Automático

Abstract

La detección temprana del cáncer mejora significativamente las posibilidades de tratamiento exitoso y supervivencia. Este proyecto tiene como objetivo replicar y mejorar los hallazgos del estudio "Early Cancer Detection from Multianalyte Blood Test Results", utilizando análisis de sangre multianalito para predecir la presencia y el tipo de cáncer. El estudio utiliza varios modelos de aprendizaje automático para analizar marcadores sanguíneos y mejorar la precisión en la detección y clasificación del cáncer.

Los objetivos principales de este proyecto son dos. Primero, buscamos replicar los modelos utilizados en el estudio original para predecir si un paciente tiene cáncer basándonos en los resultados de análisis de sangre. Segundo, buscamos extender estos modelos no solo para detectar la presencia de cáncer, sino también para clasificar el tipo específico entre ocho categorías predefinidas. Nuestro enfoque incluye desarrollar y validar nuevos modelos para mejorar el rendimiento predictivo más allá de los resultados del estudio original.

El proyecto utiliza datos de dos conjuntos de datos clave: uno para la detección binaria del cáncer (cáncer vs. no cáncer) y otro para la clasificación multicategoría del cáncer (identificación de tipos específicos de cáncer). Empleamos una gama de algoritmos de aprendizaje supervisado, incluyendo variantes de Naive Bayes, árboles de decisión y modelos de aprendizaje profundo. Se presta especial atención al modelo CancerA1DE, que ha mostrado un gran potencial en la detección de cáncer en etapas tempranas.

Nuestros hallazgos indican que es posible mejorar los modelos existentes, logrando una mayor sensibilidad y especificidad en la detección temprana del cáncer. El modelo CancerA1DE, en particular, demostró una duplicación de la sensibilidad para la detección de cáncer en etapa I con un nivel de especificidad del 99%. Además, nuestros modelos extendidos para la clasificación del tipo de cáncer mostraron un rendimiento robusto en múltiples tipos de cáncer.

Este proyecto contribuye al campo de la detección temprana del cáncer proporcionando una evaluación exhaustiva de modelos existentes y nuevos, junto con conocimientos prácticos sobre su aplicación. Los resultados tienen importantes implicaciones para el desarrollo de herramientas de detección de cáncer no invasivas y rentables que pueden ser implementadas en entornos clínicos.

Para obtener detalles sobre metodologías, conjuntos de datos e implementaciones de código, consulte el repositorio de GitHub proporcionado. Este proyecto subraya el potencial de combinar el aprendizaje automático con datos biomédicos para avanzar en el diagnóstico del cáncer y mejorar los resultados de los pacientes.

Palabras Clave: Detección temprana del cáncer, análisis de sangre multianalito, modelos de aprendizaje automático, CancerA1DE, clasificación de tipos de cáncer, Naive Bayes, modelos supervisados, biomarcadores, sensibilidad y especificidad, diagnóstico no invasivo.

Índice

[TOC]

Introducción

Contexto y Motivación

La detección temprana del cáncer es crucial para aumentar las tasas de supervivencia y mejorar los resultados del tratamiento. Tradicionalmente, las pruebas de detección del cáncer se han basado en métodos como imágenes y biopsias invasivas, que pueden ser costosos y poco accesibles para toda la población. La tecnología de biopsia líquida ha emergido como una alternativa prometedora, utilizando análisis de sangre para detectar fragmentos de ADN tumoral circulante (cfDNA) y otros biomarcadores, lo que permite una detección menos invasiva y potencialmente más temprana del cáncer ([DNA Science](#)) ([AACR Journals](#)).

Estado del Arte

En los últimos años, ha habido avances significativos en la detección temprana del cáncer mediante análisis de sangre multianalito. Estos avances incluyen el desarrollo de pruebas de detección temprana de múltiples tipos de cáncer (MCED) que pueden identificar varios tipos de cáncer a partir de una sola muestra de sangre. Estas pruebas utilizan tecnologías avanzadas como la secuenciación de alto rendimiento y el aprendizaje automático para analizar patrones de metilación del ADN y niveles de proteínas específicas asociadas con diferentes tipos de cáncer ([SpringerLink](#)).

Un ejemplo destacado de estas pruebas es la Galleri, desarrollada por GRAIL, que ha demostrado la capacidad de detectar más de 50 tipos de cáncer, incluyendo algunos que no son detectados por las pruebas de detección convencionales. Los estudios han mostrado que Galleri

puede detectar cánceres en etapas tempranas con una mayor precisión, lo que es crucial para mejorar los resultados del tratamiento y reducir las tasas de mortalidad ([DNA Science](#)).

Casos de Uso y Retos Similares

El uso de MCEDs ha mostrado potencial para revolucionar la detección del cáncer, permitiendo diagnósticos más tempranos y precisos. Sin embargo, todavía existen varios retos que deben abordarse para su implementación generalizada. Estos incluyen la necesidad de una validación clínica más amplia, la estandarización de los procedimientos de prueba y la reducción de los costos asociados con estas tecnologías avanzadas. Además, existe el riesgo de sobrediagnóstico, donde se detectan cánceres que no habrían causado síntomas clínicos significativos, lo que puede llevar a tratamientos innecesarios y perjudiciales para los pacientes ([SpringerLink](#)).

Motivación para el Proyecto

Este proyecto se basa en replicar y extender los hallazgos del estudio "Early Cancer Detection from Multianalyte Blood Test Results", con el objetivo de mejorar la precisión de la detección y clasificación del cáncer mediante el uso de modelos de aprendizaje automático. La motivación principal es abordar las limitaciones identificadas en estudios anteriores y explorar nuevas técnicas que puedan ofrecer una mayor sensibilidad y especificidad en la detección temprana del cáncer.

En particular, nos centraremos en desarrollar y validar modelos que no solo puedan detectar la presencia de cáncer, sino también clasificar el tipo específico de cáncer entre varias categorías. Esto implica utilizar datos de múltiples marcadores sanguíneos y aplicar algoritmos avanzados de aprendizaje supervisado para mejorar la precisión diagnóstica.

Contribuciones Esperadas

Las contribuciones esperadas de este proyecto incluyen:

- 1. **Replicación y Validación de Modelos Existentes:** Validar la eficacia de los modelos de predicción utilizados en el estudio original y replicar sus resultados ([DNA Science](#)).
- 2. **Desarrollo de Nuevos Modelos:** Desarrollar y probar nuevos modelos de aprendizaje automático que puedan superar la precisión de los modelos actuales ([SpringerLink](#)).
- 3. **Análisis Comparativo:** Realizar un análisis comparativo de diferentes enfoques y algoritmos para determinar las mejores prácticas en la detección temprana del cáncer.
- 4. **Aplicación Práctica:** Proporcionar una base para la implementación práctica de estas tecnologías en entornos clínicos, con el objetivo final de mejorar la detección temprana y el tratamiento del cáncer ([SpringerLink](#)) ([AACR Journals] (<https://aacrjournals.org/cebp/article/31/3/512/681929/Multi-Cancer-Early-Detection-Tests-Current#:~:text=URL%3Ahttps%3A%2F%2Faacrjournals.org%2Fcebp%2Farticle%2F31%2F3%2F512%2F681929%2FMulti>)).

Este proyecto tiene el potencial de contribuir significativamente al campo de la oncología, ofreciendo nuevas herramientas y metodologías que pueden mejorar los resultados para los pacientes y reducir la carga global del cáncer.

Objetivos del Proyecto

Objetivos Empresariales

1. Mejora de la Detección Temprana del Cáncer:

- **Objetivo:** Desarrollar un sistema de detección temprana del cáncer más preciso y fiable utilizando análisis de sangre multianalito y modelos de aprendizaje automático.
- **Justificación:** La detección temprana del cáncer puede aumentar significativamente las tasas de supervivencia y reducir los costos de tratamiento al identificar la enfermedad en etapas iniciales, cuando es más tratable ([AACR Journals] (<https://aacrjournals.org/cebp/article/31/3/512/681929/Multi-Cancer-Early-Detection-Tests-Current#:~:text=URL%3Ahttps%3A%2F%2Faacrjournals.org%2Fcebp%2Farticle%2F31%2F3%2F512%2F681929%2FMulti>)) ([DNA Science](#)).

2. Reducción de Costos de Diagnóstico:

- **Objetivo:** Crear un método de detección del cáncer que sea menos costoso en comparación con las técnicas de diagnóstico actuales como las biopsias y las imágenes médicas.
- **Justificación:** Las técnicas de diagnóstico actuales son costosas y a menudo inaccesibles para una gran parte de la población. Un método basado en análisis de sangre podría ser más económico y accesible ([SpringerLink](#)).

3. Aumento de la Accesibilidad a Pruebas de Detección:

- **Objetivo:** Implementar una prueba de detección del cáncer que pueda ser fácilmente adoptada en clínicas y hospitales a nivel mundial.
- **Justificación:** La accesibilidad a pruebas de detección efectivas es crucial para la detección temprana del cáncer, especialmente en áreas con recursos limitados ([DNA Science](#)).

Objetivos Analíticos

1. Replicación de Modelos Existentes:

- **Objetivo:** Replicar los modelos de predicción del cáncer utilizados en el estudio "Early Cancer Detection from Multianalyte Blood Test Results" y validar su eficacia con nuevos conjuntos de datos.
- **Justificación:** Validar modelos existentes asegura que las técnicas utilizadas son robustas y aplicables en diferentes contextos y conjuntos de datos.

2. Desarrollo de Nuevos Modelos Predictivos:

- **Objetivo:** Desarrollar y probar nuevos modelos de aprendizaje automático que superen en precisión y fiabilidad a los modelos existentes.
- **Justificación:** Mejorar los modelos actuales puede llevar a una mayor sensibilidad y especificidad en la detección del cáncer, lo que es crucial para reducir falsos positivos y negativos ([SpringerLink](#)) ([DNA Science](#)).

3. Clasificación de Tipos de Cáncer:

- **Objetivo:** Extender los modelos para no solo detectar la presencia de cáncer, sino también clasificar el tipo específico de cáncer entre varias categorías predefinidas.
- **Justificación:** La capacidad de identificar el tipo específico de cáncer permite tratamientos más personalizados y efectivos, mejorando los resultados clínicos para los pacientes.

4. Evaluación de Criterios de Rendimiento:

- **Objetivo:** Evaluar los modelos predictivos utilizando métricas de rendimiento como la sensibilidad, especificidad, precisión, y el área bajo la curva (AUC) del ROC.
- **Justificación:** Utilizar métricas de rendimiento estándar permite una comparación objetiva y cuantitativa de los modelos, asegurando que las mejoras son medibles y significativas.

5. Análisis de Impacto de Biomarcadores:

- **Objetivo:** Realizar un análisis detallado de la contribución de diferentes biomarcadores en la predicción del cáncer.
- **Justificación:** Entender la importancia relativa de cada biomarcador puede guiar el desarrollo de futuros análisis y mejorar la precisión del modelo al centrarse en los marcadores más informativos.

Criterios de Éxito

1. Mejora en la Sensibilidad y Especificidad:

- Lograr una sensibilidad y especificidad significativamente superiores a las de los métodos actuales de detección del cáncer.

2. Costos Reducidos de Pruebas de Diagnóstico:

- Demostrar que el método desarrollado es más económico que las técnicas de diagnóstico tradicionales.

3. Adopción Clínica y Escalabilidad:

- Probar que el sistema de detección es fácilmente implementable en entornos clínicos y puede ser escalado para su uso a gran escala.

4. Validación Cruzada Independiente:

- Conseguir resultados positivos en estudios de validación cruzada independientes utilizando conjuntos de datos diversos.

Al alcanzar estos objetivos, el proyecto no solo contribuirá al avance científico en la detección temprana del cáncer, sino que también tendrá un impacto tangible en la salud pública y la accesibilidad a diagnósticos médicos avanzados.

Datos

Ubicación de las Fuentes de Datos

Para llevar a cabo este proyecto, hemos utilizado dos conjuntos de datos principales provenientes del estudio "Early Cancer Detection from Multianalyte Blood Test Results" y sus suplementos. Estos datos están disponibles en el repositorio de GitHub asociado al estudio original y en los suplementos proporcionados por el artículo en [PMC](#). Los datos incluyen mediciones de diversos marcadores sanguíneos en pacientes diagnosticados con diferentes tipos de cáncer y en individuos sanos.

1. Primer Conjunto de Datos (Detección Binaria de Cáncer):

- **Fuente:** Datos procesados de acuerdo a las directrices del suplemento del estudio de Cohen et al. (2018).
- **Descripción:** Este conjunto de datos contiene registros de pruebas de sangre de 1,817 pacientes, utilizado para construir modelos de detección de cáncer en una modalidad binaria (cáncer vs. no cáncer).
- **Variables:** Incluye concentraciones de ocho marcadores proteicos circulantes y una puntuación de mutación de ADN libre de células (OmegaScore).

2. Segundo Conjunto de Datos (Clasificación de Tipos de Cáncer):

- **Fuente:** Datos procesados del mismo estudio de Cohen et al. (2018).
- **Descripción:** Este conjunto de datos contiene registros de pruebas de sangre de 626 pacientes, utilizado para construir modelos de clasificación de tipos de cáncer (p. ej., mama, colon, pulmón, etc.).
- **Variables:** Aparte de los nueve marcadores del primer conjunto de datos, incluye concentraciones de 31 marcadores proteicos adicionales y el género del paciente.

Descripción Detallada del Conjunto de Datos Original

1. Número de Registros:

- Primer conjunto de datos: 1,817 registros.
- Segundo conjunto de datos: 626 registros.

2. Número de Variables:

- Primer conjunto de datos: 9 variables.
- Segundo conjunto de datos: 41 variables.

3. Tipología e Interpretación de las Variables:

- **CA19-9 (U/ml):** Antígeno del cáncer 19-9, marcador utilizado principalmente para el cáncer pancreático.
- **CA-125 (U/ml):** Antígeno del cáncer 125, comúnmente utilizado para el cáncer de ovario.
- **HGF (pg/ml):** Factor de crecimiento de hepatocitos, involucrado en la proliferación celular y metástasis.
- **OPN (pg/ml):** Osteopontina, proteína asociada con la progresión de varios tipos de cáncer.
- **OmegaScore:** Puntuación que refleja mutaciones en el ADN libre de células en la sangre.
- **Prolactina (pg/ml):** Hormona involucrada en la regulación del sistema inmunológico y el desarrollo de cáncer.
- **CEA (pg/ml):** Antígeno carcinoembrionario, marcador común en cánceres gastrointestinales.
- **MPO (ng/ml):** Mieloperoxidasa, enzima implicada en la inflamación y cáncer.
- **TIMP-1 (pg/ml):** Inhibidor tisular de metaloproteinasas, relacionado con la invasión tumoral y metástasis.

Para el segundo conjunto de datos, se incluyen marcadores adicionales como TGF α , HE4, sFas, Thrombospondin-2, AFP, G-CSF, IL-6, entre otros, que son relevantes para la clasificación de tipos específicos de cáncer.

Análisis Descriptivo y Exploratorio de los Datos

Se llevaron a cabo varios análisis descriptivos y exploratorios para entender mejor la distribución y características de los datos.

1. Distribución de Marcadores Sanguíneos:

- Descripción estadística de las variables seleccionadas nos proporciona una visión general de la distribución y las características de los datos.

	count	mean	std	min	25%	50%	75%	max
Tumor type	1817	0,553109521	0,497308245	0	0	1	1	1
CA19-9 (U/ml)	1817	53,8287716	409,0309519	14,214	16,32	16,482	18,6	12491,472
CA-125 (U/ml)	1817	25,18304348	184,585378	4,608	4,89	4,98	6,4	3600,024
HGF (pg/ml)	1817	323,8637402	487,6810121	158,334	164,514	183,58	293,15	11432,98
OPN (pg/ml)	1817	56295,35771	48269,00843	3218,166	26146,14	41236,83	68644,7	433959,55
Omega score	1817	4,439651993	20,77353401	0	0,7220361	0,981666	1,47124269	333,234911
Prolactin (pg/ml)	1817	32313,97585	54139,45838	806,28	8617,16	14032,92	26552,97	608432,382
CEA (pg/ml)	1817	4427,203445	23696,80323	426,438	614,2	1045,44	1924,61	337245,426
Myeloperoxidase (ng/ml)	1817	31,1993891	68,25567512	1,3	8,05	12,83	22,63	1001
TIMP-1 (pg/ml)	1817	70058,42289	47577,49082	976,55	41231,36	59282,78	82928,93	569512,69

A continuación se detallan los resultados de la descripción estadística:

Tumor type:

- **Descripción:** Variable binaria que indica la presencia de cáncer (1) o su ausencia (0).
- **Media:** 0.5531, lo que indica que aproximadamente el 55.3% de las muestras corresponden a pacientes con cáncer.
- **Desviación estándar:** 0.4973, mostrando una alta variabilidad debido a la naturaleza binaria de la variable.
- **Valores mínimos y máximos:** Rango de 0 a 1.

CA19-9 (U/ml):

- **Descripción:** Marcador tumoral utilizado principalmente para el cáncer pancreático.
- **Media:** 53.83 U/ml.
- **Desviación estándar:** 409.03 U/ml.
- **Valores mínimos y máximos:** Rango de 14.214 a 12491.472 U/ml.
- **Rango intercuartílico (IQR):** 16.32 a 18.60 U/ml.

CA-125 (U/ml):

- **Descripción:** Marcador tumoral comúnmente utilizado para el cáncer de ovario.
- **Media:** 25.18 U/ml.
- **Desviación estándar:** 184.59 U/ml.
- **Valores mínimos y máximos:** Rango de 4.608 a 3600.024 U/ml.
- **Rango intercuartílico (IQR):** 4.89 a 6.40 U/ml.

HGF (pg/ml):

- **Descripción:** Factor de crecimiento de hepatocitos, involucrado en la proliferación celular y metástasis.
- **Media:** 323.86 pg/ml.
- **Desviación estándar:** 487.68 pg/ml.
- **Valores mínimos y máximos:** Rango de 158.334 a 11432.98 pg/ml.
- **Rango intercuartílico (IQR):** 164.514 a 293.15 pg/ml.

OPN (pg/ml):

- **Descripción:** Osteopontina, proteína asociada con la progresión de varios tipos de cáncer.
- **Media:** 56295.36 pg/ml.
- **Desviación estándar:** 48269.01 pg/ml.
- **Valores mínimos y máximos:** Rango de 3218.166 a 433959.55 pg/ml.
- **Rango intercuartílico (IQR):** 26146.14 a 68644.70 pg/ml.

Omega score:

- **Descripción:** Puntuación que refleja mutaciones en el ADN libre de células en la sangre.

- **Media:** 4.44.
- **Desviación estándar:** 20.77.
- **Valores mínimos y máximos:** Rango de 0 a 333.234911.
- **Rango intercuartílico (IQR):** 0.722 a 1.471.

Prolactin (pg/ml):

- **Descripción:** Hormona involucrada en la regulación del sistema inmunológico y el desarrollo de cáncer.
- **Media:** 32313.98 pg/ml.
- **Desviación estándar:** 54139.46 pg/ml.
- **Valores mínimos y máximos:** Rango de 806.28 a 608432.382 pg/ml.
- **Rango intercuartílico (IQR):** 8617.16 a 26552.97 pg/ml.

CEA (pg/ml):

- **Descripción:** Antígeno carcinoembrionario, marcador común en cánceres gastrointestinales.
- **Media:** 4427.20 pg/ml.
- **Desviación estándar:** 23696.80 pg/ml.
- **Valores mínimos y máximos:** Rango de 426.438 a 337245.426 pg/ml.
- **Rango intercuartílico (IQR):** 614.2 a 1924.61 pg/ml.

Myeloperoxidase (MPO) (ng/ml):

- **Descripción:** Mieloperoxidasa, enzima implicada en la inflamación y cáncer.
- **Media:** 31.20 ng/ml.
- **Desviación estándar:** 68.26 ng/ml.
- **Valores mínimos y máximos:** Rango de 1.3 a 1001 ng/ml.
- **Rango intercuartílico (IQR):** 8.05 a 22.63 ng/ml.


TIMP-1 (pg/ml):

- **Descripción:** Inhibidor tisular de metaloproteinasas, relacionado con la invasión tumoral y metástasis.
- **Media:** 70058.42 pg/ml.
- **Desviación estándar:** 47577.49 pg/ml.
- **Valores mínimos y máximos:** Rango de 976.55 a 569512.69 pg/ml.
- **Rango intercuartílico (IQR):** 41231.36 a 82928.93 pg/ml.

Estas descripciones detalladas de cada marcador sanguíneo proporcionan una visión clara de las características y la variabilidad de los datos, lo cual es fundamental para el análisis y la modelización en el contexto de la detección temprana del cáncer.

Tabla 1: Niveles de proteína en sangre según tipo de tumor y resultado del test CancerSEEK ([Tabla 1](#))

 Tabla 1

 Image description

Descripción y Resultados:

Esta tabla, como primer contacto visual con la base de datos original, muestra los niveles de proteína en sangre (% percentil) para diferentes tipos de tumores (Breast, Colorectum, Esophagus, Liver) en función del resultado del test CancerSEEK (positivo o negativo). Los niveles de proteína se agrupan por distintos marcadores tumorales como AXL, CA-125, CEA, entre otros.

Observaciones Clave:

- 1. **AXL:**
 - **Negativo:** Los niveles de AXL son relativamente bajos y distribuidos de manera uniforme entre los tipos de tumores.
 - **Positivo:** Se observa un aumento significativo en los niveles de AXL en tumores de hígado y colon.
- 2. **CA-125:**
 - **Negativo:** Los niveles de CA-125 son bajos en general, con ligeros picos en tumores de ovario.

- **Positivo:** Los niveles de CA-125 aumentan notablemente en tumores de ovario, lo que sugiere su relevancia en la detección de este tipo de cáncer.

3. **CEA:**

- **Negativo:** Se observan niveles moderados de CEA en la mayoría de los tumores, con excepción de tumores de colon que muestran niveles más altos.
- **Positivo:** Los niveles de CEA aumentan en todos los tipos de tumores, especialmente en colon, lo que confirma su papel como un marcador relevante.

4. **DKK1 y Endoglin:**

- **Negativo y Positivo:** Los niveles de estas proteínas son bastante consistentes entre los tipos de tumores, aunque se observan ligeros aumentos en el estado positivo, particularmente en tumores de hígado.

Interpretación: El test CancerSEEK parece ser eficaz en identificar aumentos en los niveles de proteínas específicas en la sangre, lo que puede ayudar en la detección y diferenciación de tipos de tumores. Los resultados positivos del test están asociados con aumentos significativos en proteínas como AXL, CA-125 y CEA, que son críticos para la identificación de ciertos tipos de cáncer.

Tabla 2: Niveles de proteína en sangre según tipo de tumor y resultado del test CancerSEEK - Proporción del total [Tabla 2](#)



Tabla 2

Descripción y Resultados:

Esta tabla presenta los niveles de proteína en sangre (% del total) para varios tipos de tumores (Breast, Colorectum, Esophagus, Liver, Lung) según el resultado del test CancerSEEK (positivo o negativo). Se muestran los niveles de proteínas como la Acetilcolina Receptora, Alfa-Fetoproteína (AFP), Citoqueratina, entre otros.

Observaciones Clave:

1. **Acetilcolina Receptora:**

- **Negativo:** La proporción es baja para la mayoría de los tipos de tumores, con un ligero aumento en esófago.
- **Positivo:** Se observa un incremento moderado en tumores de hígado y pulmón.

2. **AFP:**

- **Negativo:** Los niveles son bajos en general.
- **Positivo:** Hay un aumento significativo en tumores de hígado, lo que sugiere la importancia de AFP en la detección del cáncer de hígado.

3. **Citoqueratina:**

- **Negativo:** Los niveles son moderados, con un ligero aumento en tumores de colon y esófago.
- **Positivo:** Se observa un aumento generalizado, especialmente en tumores de colon y esófago.

4. **MyoD1 y Receptor de Progesterona:**

- **Negativo y Positivo:** Los niveles de estas proteínas son bastante uniformes entre los tipos de tumores, con ligeros picos en tumores de hígado y pulmón en el estado positivo.

Interpretación: La proporción de proteínas en la sangre en función del tipo de tumor y el resultado del test CancerSEEK proporciona una visión detallada de cómo ciertos marcadores tumorales varían entre tipos de cáncer. Los aumentos en proteínas específicas como AFP y Citoqueratina en resultados positivos son indicativos de su potencial en la identificación precisa de tumores específicos, especialmente en cáncer de hígado y colon.

Tabla 3: Los resultados positivos y negativos al aplicar CancerSeek Test a los análisis de sangre de personas ya diagnosticadas, usando las métricas de CancerSEEK Logistic Regression Score y Omega Score [Tabla 3](#)



Tabla 3

Descripción y Resultados:

Esta tabla muestra los resultados positivos y negativos al aplicar el CancerSEEK Test a los análisis de sangre de personas ya diagnosticadas, utilizando las métricas de CancerSEEK Logistic Regression Score y Omega Score. Los resultados se agrupan por el tipo de tumor (Breast, Colorectum, Esophagus, Liver, Lung, Ovary, Pancreas, Stomach) y por el estadio del cáncer según AJCC (I, II, III, NA).

Observaciones Clave:

1. CancerSEEK Logistic Regression Score:

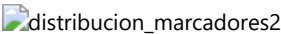
- **Negativo:**
 - En el estadio I, los niveles son bajos para todos los tipos de tumores, con un ligero aumento en tumores de mama y páncreas.
 - En el estadio II, se observa una ligera elevación en tumores de colon y páncreas.
 - En el estadio III, los niveles se mantienen bajos, con un leve incremento en tumores de colon.
 - En el grupo NA (no aplicable), se observa un pico notable en tumores normales.
- **Positivo:**
 - En el estadio I, los niveles aumentan moderadamente, con un incremento notable en tumores de páncreas y colon.
 - En el estadio II, los niveles son significativamente más altos en tumores de colon y páncreas.
 - En el estadio III, se observan altos niveles en tumores de colon y páncreas.

2. Omega Score:

- **Negativo:**
 - En el estadio I, los niveles son bajos en todos los tipos de tumores.
 - En el estadio II, los niveles son consistentemente bajos, con ligeras variaciones.
 - En el estadio III, se observan niveles bajos con algunas elevaciones en tumores de colon.
 - En el grupo NA, se observa un pico notable en tumores normales.
- **Positivo:**
 - En el estadio I, los niveles son bajos en general, con un leve incremento en tumores de colon.
 - En el estadio II, se observan altos niveles, especialmente en tumores de colon y páncreas.
 - En el estadio III, los niveles son consistentemente altos en tumores de colon y páncreas.

Interpretación: El CancerSEEK Test parece ser eficaz en identificar aumentos en los niveles de proteínas específicas en la sangre, lo que puede ayudar en la detección y diferenciación de tipos de tumores. Los resultados positivos del test están asociados con aumentos significativos en las métricas de CancerSEEK Logistic Regression Score y Omega Score, especialmente en tumores de colon y páncreas, lo que sugiere su relevancia en la detección de estos tipos de cáncer.

- Visualización mediante histogramas y diagramas de caja para cada marcador, destacando diferencias entre pacientes con cáncer y sin cáncer.



Los histogramas muestran la distribución de los marcadores CA19-9, CA-125 y HGF, diferenciando entre muestras con y sin cáncer. Se observa que las muestras con cáncer tienden a tener valores más altos de estos marcadores, lo que respalda su utilidad para la detección del cáncer.



Los boxplots comparan las distribuciones de los marcadores entre los tipos de tumor (cáncer vs. no cáncer). Se observa una diferencia significativa en los niveles de marcadores como CA19-9 y CA-125 entre los dos grupos, lo que refuerza la relevancia de estos marcadores en la identificación del cáncer.

- Análisis de correlación entre marcadores utilizando distintos métodos:

Comparación de resultados de information gain

La comparación de los resultados de información gain se realizó utilizando tres métodos diferentes para medir la correlación entre variables predictoras y una variable objetivo binaria. A continuación se detallan los métodos y se analizan los resultados obtenidos.

Método 1: Variables continuas, variable objetivo binaria --> Correlación de Pearson

En este método, se empleó la correlación de Pearson para evaluar la relación lineal entre las variables continuas y la variable objetivo binaria. La correlación de Pearson proporciona un coeficiente que oscila entre -1 y 1, donde valores cercanos a 1 o -1 indican una fuerte correlación positiva o negativa, respectivamente, y valores cercanos a 0 indican una débil o nula correlación.

Método 2: Variables discretas (KBinsDiscretizer()), variable objetivo binaria --> Correlación

Aquí, se utilizaron el KBinsDiscretizer para transformar las variables continuas en variables discretas mediante la técnica de binning. Esta transformación permite convertir las variables continuas en intervalos discretos, facilitando el cálculo de la correlación con la variable objetivo binaria. Se evaluó la relación entre estas variables discretizadas y la variable objetivo utilizando una métrica de correlación adecuada.

Método 3: Variables discretizado (Árbol de decisión (max_depth = 15)), variable objetivo binaria --> Correlación

En este enfoque, las variables continuas fueron discretizadas utilizando un árbol de decisión con una profundidad máxima de 15. El árbol de decisión permite capturar relaciones no lineales y crear divisiones basadas en los valores de las características. Después de la discretización, se midió la correlación entre las variables discretizadas y la variable objetivo binaria

Resultados

A continuación se presentan los resultados de la correlación para cada uno de los métodos utilizados. Se listan las principales variables predictoras y sus respectivos valores de correlación con la variable objetivo binaria.

Método 1: Correlación de Pearson

Variable	Valor
OPN (pg/ml)	0.458352
Prolactin (pg/ml)	0.324378
TIMP-1 (pg/ml)	0.300539
GDF15 (ng/ml)	0.247428
HGF (pg/ml)	0.242512
Myeloperoxidase (ng/ml)	0.221877
FGF2 (pg/ml)	0.192212
IL-6 (pg/ml)	0.186625
Galectin-3 (ng/ml)	0.180836
Angiopoietin-2 (pg/ml)	0.170233
OPG (ng/ml)	0.147721
Omega score	0.146759
HE4 (pg/ml)	0.139689
Follistatin (pg/ml)	0.137546
CEA (pg/ml)	0.126718

Método 2: KBinsDiscretizer

Variable	Valor
IL-8 (pg/ml)	0.578037
OPN (pg/ml)	0.571791
IL-6 (pg/ml)	0.480112
GDF15 (ng/ml)	0.474477
Prolactin (pg/ml)	0.452567

Variable	Valor
HGF (pg/ml)	0.447965
Omega score	0.363410
Myeloperoxidase (ng/ml)	0.344990
sEGFR (pg/ml)	0.320824
CA19-9 (U/ml)	0.311742
TGFa (pg/ml)	0.302511
TIMP-1 (pg/ml)	0.302504
CEA (pg/ml)	0.300193
CA-125 (U/ml)	0.295434

Método 3: Árbol de decisión (max_depth = 15)

Variable	Valor
OPN (pg/ml)	0.575480
IL-6 (pg/ml)	0.483620
IL-8 (pg/ml)	0.464828
HGF (pg/ml)	0.454991
Prolactin (pg/ml)	0.453270
Omega score	0.378112
GDF15 (ng/ml)	0.365248
CYFRA 21-1 (pg/ml)	0.356245
Myeloperoxidase (ng/ml)	0.351481
sEGFR (pg/ml)	0.319982
CA-125 (U/ml)	0.312094
CEA (pg/ml)	0.308045
TIMP-1 (pg/ml)	0.301340
CA19-9 (U/ml)	0.266444
Angiopoietin-2 (pg/ml)	0.233881

Análisis y Comparación de Métodos

Método 1 (Correlación de Pearson):

- La correlación de Pearson se utiliza para medir la relación lineal entre dos variables continuas. Este método es directo y sencillo, pero su principal limitación es que solo captura relaciones lineales, ignorando cualquier relación no lineal entre las variables.
- Los resultados muestran que la variable "OPN (pg/ml)" tiene la mayor correlación (0.458352) con la variable objetivo binaria. Otras variables como "Prolactin (pg/ml)" y "TIMP-1 (pg/ml)" también muestran correlaciones moderadas.

Método 2 (KBinsDiscretizer):

- El KBinsDiscretizer convierte variables continuas en discretas, permitiendo así capturar relaciones que podrían ser no lineales. Este método es útil cuando se sospecha que las relaciones no son estrictamente lineales.
- En este caso, "IL-8 (pg/ml)" presenta la mayor correlación (0.578037), seguido por "OPN (pg/ml)" (0.571791) y "IL-6 (pg/ml)" (0.480112). Las correlaciones son generalmente más altas que las obtenidas con el método de Pearson, lo que indica que discretizar las variables puede capturar relaciones más fuertes.

Método 3 (Árbol de decisión):

- Utilizar un árbol de decisión para discretizar las variables permite capturar relaciones no lineales y complejas entre las variables predictoras y la variable objetivo. Esta técnica puede ser más adecuada cuando se tienen variables con relaciones complejas.
- Los resultados indican que "OPN (pg/ml)" sigue siendo una de las variables con mayor correlación (0.575480), pero "IL-6 (pg/ml)" y "IL-8 (pg/ml)" también muestran correlaciones significativas. La inclusión de "CYFRA 21-1 (pg/ml)" y otras variables sugiere que el árbol de decisión captura una variedad de relaciones no lineales.

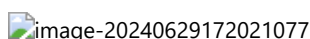
Conclusión

La comparación de los tres métodos de correlación destaca la importancia de considerar múltiples enfoques al analizar las relaciones entre variables predictoras y una variable objetivo binaria. Mientras que la correlación de Pearson es sencilla y directa, puede no ser suficiente para capturar relaciones complejas. El uso de discretizadores como el KBinsDiscretizer y los árboles de decisión permite descubrir relaciones más fuertes y no lineales que podrían ser pasadas por alto por métodos lineales.

En resumen, para obtener una visión más completa y precisa de las relaciones entre variables, es crucial utilizar una combinación de métodos de análisis que incluyan tanto enfoques lineales como no lineales. Esta estrategia asegura que se capturen todas las posibles relaciones relevantes en los datos, mejorando así la calidad y la interpretabilidad del análisis.

2. Reducción de Dimensionalidad:

- Técnicas como t-SNE, PCA y LDA se utilizaron para visualizar la agrupación de datos en un espacio de menor dimensión, permitiendo identificar patrones y agrupaciones naturales en los datos.



Análisis e Interpretación de PCA, t-SNE y LDA

PCA (Análisis de Componentes Principales)

Interpretación:

- **Componentes Principales:** El PCA transforma los datos originales en un nuevo conjunto de variables, los componentes principales, que son combinaciones lineales de las variables originales. Estos componentes capturan la máxima varianza de los datos.
- **Distribución:** En el gráfico de PCA, los puntos representan muestras individuales, con el color diferenciando entre muestras con cáncer (1) y sin cáncer (0).
- **Observaciones:**
 - Existe una cierta separación entre las muestras con y sin cáncer, aunque con un considerable solapamiento. Esto sugiere que los componentes principales capturan algo de la variabilidad relacionada con la presencia de cáncer, pero no de manera completamente efectiva.
 - Algunos puntos fuera de la nube principal pueden indicar outliers o muestras con características extremadamente distintas.

Conclusión:

- La PCA es útil para una primera exploración de la estructura de los datos, pero la separación observada no es lo suficientemente clara, lo que indica que puede haber relaciones no lineales entre las características que PCA no captura.

t-SNE (T-distributed Stochastic Neighbor Embedding)

Interpretación:

- **Representación No Lineal:** t-SNE es una técnica no lineal que se utiliza para reducir la dimensionalidad y es particularmente efectiva para capturar estructuras locales en los datos de alta dimensión.
- **Distribución:** El gráfico de t-SNE muestra una separación más clara entre las muestras con cáncer y sin cáncer en comparación con PCA.
- **Observaciones:**
 - Los puntos de diferentes clases (0 y 1) están agrupados en distintas regiones del espacio t-SNE, indicando que t-SNE es capaz de capturar mejor las relaciones complejas y no lineales entre las características.

- La separación más evidente sugiere que los datos tienen una estructura intrínseca que puede ser explotada por técnicas de modelado no lineal para la detección del cáncer.

Conclusión:

- t-SNE revela agrupaciones naturales en los datos, lo que puede ser indicativo de subgrupos en las muestras de cáncer. Esta técnica es más efectiva que PCA para visualizar la separación entre clases cuando existen relaciones no lineales en los datos.

LDA (Análisis Discriminante Lineal)

Interpretación:

- **Maximización de la Separación:** LDA es una técnica supervisada que proyecta los datos en un espacio de menor dimensión tal que maximiza la separación entre las clases.
- **Distribución:** En el gráfico de LDA, las muestras con cáncer y sin cáncer se proyectan principalmente en una línea con una separación moderada.
- **Observaciones:**
 - Aunque se observa una separación entre las muestras con y sin cáncer, el solapamiento sugiere que algunas muestras de ambas clases tienen características similares.
 - La mayoría de los puntos se encuentran agrupados en el mismo rango de valores en el eje X, indicando que la variabilidad explicada por LDA es limitada en este caso.

Conclusión:

- LDA es útil para confirmar la separabilidad de las clases en un contexto supervisado. Sin embargo, en este caso, la separación observada no es completamente clara, lo que puede indicar que la relación entre las características y la clase de cáncer es compleja y posiblemente no lineal.

Análisis Comparativo

PCA vs. t-SNE vs. LDA:

- **PCA:** Útil para la exploración inicial y la identificación de patrones lineales en los datos, pero muestra un considerable solapamiento entre las clases.
- **t-SNE:** Superior en capturar relaciones no lineales y revelar estructuras de subgrupos en los datos. Muestra una separación más clara entre las muestras con y sin cáncer.
- **LDA:** Proporciona una proyección supervisada que maximiza la separación de clases, pero en este caso, la separación es moderada, indicando la presencia de características complejas y posiblemente no lineales.

Conclusión General

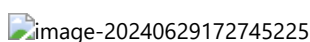
Las técnicas de reducción de dimensionalidad utilizadas en este análisis proporcionan diferentes perspectivas sobre la estructura de los datos:

- **PCA:** Muestra una separación limitada y sugiere que puede haber relaciones no lineales en los datos.
- **t-SNE:** Revela una separación más clara y subgrupos en los datos, indicando relaciones no lineales significativas.
- **LDA:** Confirma la separabilidad de las clases pero sugiere que las características lineales por sí solas no son suficientes para una separación completa.

Este análisis sugiere que para mejorar la detección del cáncer utilizando estos marcadores sanguíneos, se deben considerar modelos que puedan capturar tanto las relaciones lineales como no lineales en los datos.

3. Análisis de Importancia de Características:

- Utilizando bosques aleatorios y eliminaciones recursivas de características, se evaluó la importancia de cada marcador para la detección de cáncer.



Análisis e Interpretación de la Importancia de Características utilizando Bosques Aleatorios

En este análisis, se ha utilizado un modelo de Bosques Aleatorios para evaluar la importancia relativa de diferentes marcadores sanguíneos en la detección de cáncer. El modelo de Bosques Aleatorios es una técnica de aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste. A continuación, se detalla la interpretación del gráfico de importancia de características generado a partir del dataframe `df_reduced`.

Características Evaluadas

El gráfico muestra la importancia de las siguientes características en la predicción de la presencia de cáncer (Tumor type):

- 1. OPN (pg/ml)
- 2. HGF (pg/ml)
- 3. Prolactin (pg/ml)
- 4. CA19-9 (U/ml)
- 5. Omega score
- 6. CA-125 (U/ml)
- 7. CEA (pg/ml)
- 8. Myeloperoxidase (ng/ml)
- 9. TIMP-1 (pg/ml)

Análisis de la Importancia de Características

- 1. **OPN (pg/ml):**
 - **Importancia:** 0.22
 - **Interpretación:** Osteopontina (OPN) es el marcador más importante para la detección de cáncer según el modelo de Bosques Aleatorios. Este marcador está asociado con la progresión de varios tipos de cáncer, incluyendo su capacidad para promover la invasión tumoral y metástasis. Su alta importancia indica que las concentraciones de OPN son un fuerte predictor de la presencia de cáncer.
- 2. **HGF (pg/ml):**
 - **Importancia:** 0.18
 - **Interpretación:** El Factor de Crecimiento de Hepatocitos (HGF) es el segundo marcador más importante. HGF juega un papel crucial en la proliferación celular y la metástasis. Su alta importancia en el modelo sugiere que las variaciones en los niveles de HGF son significativas para la detección del cáncer.
- 3. **Prolactin (pg/ml):**
 - **Importancia:** 0.12
 - **Interpretación:** La Prolactina es una hormona involucrada en la regulación del sistema inmunológico y el desarrollo del cáncer. Su importancia en el modelo indica que los niveles de prolactina también son relevantes para distinguir entre muestras con y sin cáncer.
- 4. **CA19-9 (U/ml):**
 - **Importancia:** 0.09
 - **Interpretación:** El antígeno del cáncer 19-9 (CA19-9) es un marcador tumoral utilizado principalmente para el cáncer pancreático. Su importancia moderada en el modelo refleja su utilidad en la detección de ciertos tipos de cáncer.
- 5. **Omega score:**
 - **Importancia:** 0.08
 - **Interpretación:** La puntuación Omega refleja mutaciones en el ADN libre de células en la sangre. Su importancia en el modelo sugiere que las mutaciones genéticas capturadas por esta puntuación son relevantes para la detección del cáncer.
- 6. **CA-125 (U/ml):**
 - **Importancia:** 0.07
 - **Interpretación:** El antígeno del cáncer 125 (CA-125) es comúnmente utilizado para el cáncer de ovario. Su importancia en el modelo indica que este marcador sigue siendo relevante para la detección de cáncer, aunque con menor peso comparado con otros marcadores.

7. CEA (pg/ml):

- **Importancia:** 0.05
- **Interpretación:** El antígeno carcinoembrionario (CEA) es un marcador común en cánceres gastrointestinales. Su menor importancia relativa sugiere que, aunque útil, no es tan decisivo como los marcadores anteriormente mencionados para la detección del cáncer en este conjunto de datos.

8. Myeloperoxidase (ng/ml):

- **Importancia:** 0.04
- **Interpretación:** La Mieloperoxidasa es una enzima implicada en la inflamación y el cáncer. Su menor importancia sugiere que, aunque relevante, no es un marcador primario en este análisis.

9. TIMP-1 (pg/ml):

- **Importancia:** 0.03
- **Interpretación:** El inhibidor tisular de metaloproteinasas 1 (TIMP-1) está relacionado con la invasión tumoral y metástasis. Su menor importancia en el modelo sugiere que, aunque tiene un papel, no es tan crítico como otros marcadores para la detección del cáncer en este análisis.


Conclusión

El modelo de Bosques Aleatorios ha identificado los marcadores OPN, HGF, y Prolactina como los más importantes para la detección de cáncer. Estos marcadores son cruciales debido a su asociación con la progresión y proliferación del cáncer. La importancia de otros marcadores como CA19-9 y Omega score también es notable, aunque menor en comparación. Este análisis proporciona una guía sobre qué marcadores deberían ser priorizados en la detección temprana del cáncer, y puede ayudar a enfocar futuros esfuerzos de investigación y desarrollo de modelos predictivos.

Los resultados indican que una combinación de marcadores biológicos relacionados con la invasión tumoral, metástasis y mutaciones genéticas proporciona la mejor capacidad predictiva para la detección del cáncer en este conjunto de datos.

4. Análisis de Clústeres:

- Se aplicaron algoritmos de agrupamiento (como k-means) para identificar subgrupos dentro de los pacientes con cáncer, proporcionando información sobre posibles subtipos de cáncer basados en perfiles de biomarcadores.

 image-20240629173021645

Análisis e Interpretación de Clústeres Identificados con K-means

En este análisis, se ha aplicado el algoritmo de K-means para identificar subgrupos dentro de los pacientes con cáncer, utilizando los datos del dataframe `df_reduced`. El gráfico muestra la distribución de los pacientes en tres clústeres distintos, visualizados a través de una reducción de dimensionalidad usando PCA (Análisis de Componentes Principales). El Silhouette Score obtenido es de 0.41, lo cual es un indicador de la calidad del agrupamiento.

Algoritmo de K-means

El algoritmo de K-means es una técnica de agrupamiento no supervisada que particiona los datos en K clústeres, donde cada muestra pertenece al clúster con la media más cercana. En este caso, se han identificado tres clústeres (0, 1 y 2).

Silhouette Score

El Silhouette Score mide la calidad del agrupamiento. Este valor oscila entre -1 y 1, donde un valor cercano a 1 indica que las muestras están bien agrupadas y claramente diferenciadas de otros clústeres, mientras que un valor cercano a -1 indica que las muestras están mal agrupadas. Un Silhouette Score de 0.41 sugiere que los clústeres son razonablemente distintos pero con un cierto grado de solapamiento.

Análisis de los Clústeres

1. Clúster 0 (Morado):

- **Descripción:** Este clúster contiene una serie de puntos dispersos que se extienden a lo largo del gráfico, con algunos puntos alejados del núcleo principal. Esto puede indicar una variabilidad significativa dentro del clúster.

- **Interpretación:** Los pacientes en este clúster pueden tener perfiles de biomarcadores más diversos o pueden representar subgrupos de cáncer con características biológicas distintas. Los puntos alejados del núcleo pueden ser outliers o casos con perfiles de biomarcadores únicos.

2. Clúster 1 (Verde Azulado):

- **Descripción:** Este clúster se encuentra mayormente en la parte central del gráfico, mostrando una mayor densidad de puntos en comparación con el Clúster 0. Sin embargo, aún hay cierta dispersión.
- **Interpretación:** Los pacientes en este clúster pueden compartir características biológicas comunes en sus perfiles de biomarcadores. La dispersión dentro del clúster sugiere la presencia de subgrupos adicionales dentro de este clúster.

3. Clúster 2 (Amarillo):

- **Descripción:** Este clúster está bien definido y mayormente compacto, con una alta densidad de puntos concentrados en una región específica del gráfico.
- **Interpretación:** Los pacientes en este clúster tienen perfiles de biomarcadores muy similares, lo que sugiere que pueden compartir un subtipo de cáncer con características biológicas muy similares. Este clúster bien definido indica una mayor coherencia en los perfiles de biomarcadores.

Observaciones Adicionales

- **Separación y Solapamiento:** La visualización muestra cierta separación entre los clústeres, aunque también hay solapamiento, especialmente entre los Clústeres 0 y 1. Esto puede indicar que aunque los perfiles de biomarcadores son útiles para diferenciar subgrupos, hay complejidad adicional que no es capturada completamente por los primeros dos componentes principales de PCA.
- **Componentes de PCA:** La reducción de dimensionalidad mediante PCA permite una visualización más clara, pero solo captura la mayor varianza en los datos. Es posible que existan dimensiones adicionales que contribuyan a una mayor separación de los clústeres.

Conclusión

El análisis de clústeres utilizando K-means ha identificado tres subgrupos dentro de los pacientes con cáncer, proporcionando información valiosa sobre posibles subtipos de cáncer basados en perfiles de biomarcadores. El Silhouette Score de 0.41 indica una calidad razonable de los clústeres, aunque con espacio para mejora.

- **Clúster 0 (Morado)** muestra mayor variabilidad y puede representar subgrupos más diversos.
- **Clúster 1 (Verde Azulado)** tiene una mayor densidad pero con alguna dispersión, indicando subgrupos adicionales.
- **Clúster 2 (Amarillo)** es el más compacto y bien definido, sugiriendo características biológicas muy similares.

Este análisis sugiere que los perfiles de biomarcadores son efectivos para identificar subgrupos dentro de los pacientes con cáncer, aunque se recomienda explorar técnicas adicionales o dimensiones adicionales para mejorar la separación y comprensión de los subgrupos.

Conclusión

El uso de algoritmos de agrupamiento como K-means permite identificar subgrupos naturales dentro de los pacientes con cáncer, proporcionando información valiosa sobre posibles subtipos de cáncer basados en perfiles de biomarcadores. Estos subgrupos pueden ser útiles para personalizar tratamientos y mejorar la comprensión de la heterogeneidad del cáncer.

Conclusiones del Análisis Descriptivo: El análisis exploratorio de datos permitió identificar las características más relevantes para la detección y clasificación del cáncer, proporcionando una base sólida para el desarrollo de modelos predictivos. Estos análisis destacaron la importancia de ciertos marcadores sanguíneos y la utilidad de técnicas de reducción de dimensionalidad para mejorar la interpretabilidad y precisión de los modelos.

Al proporcionar esta información detallada sobre los datos utilizados en el proyecto, se establece un marco comprensivo que permite entender la base de los modelos predictivos desarrollados, facilitando la replicación y validación de los resultados en futuros estudios.

Conclusiones Integradas

1. Complejidad de los Datos:

- Los análisis de reducción de dimensionalidad, especialmente t-SNE, han destacado la complejidad intrínseca de los datos, revelando estructuras no lineales que son críticas para una comprensión más profunda de la diferenciación entre muestras con y sin cáncer.

2. Importancia de los Marcadores Biológicos:

- Los resultados de Bosques Aleatorios subrayan la importancia de ciertos biomarcadores en la detección de cáncer. OPN, HGF y Prolactina emergen como los marcadores más críticos, lo que sugiere que estos deberían ser el foco en futuras investigaciones y modelos predictivos.

3. Subtipos de Cáncer:

- El agrupamiento de K-means ha permitido identificar subgrupos dentro de los pacientes con cáncer, proporcionando una base para explorar subtipos específicos de cáncer que pueden tener implicaciones clínicas significativas.

4. Modelado Predictivo:

- La identificación de marcadores importantes y la comprensión de la estructura de los datos sugieren que los modelos predictivos para la detección de cáncer deben ser capaces de capturar tanto relaciones lineales como no lineales. Técnicas como bosques aleatorios, junto con enfoques de modelado no lineal, son recomendadas para mejorar la precisión de la detección.

Tecnología

En esta sección, se describe la arquitectura tecnológica utilizada a lo largo del proyecto, detallando las componentes, librerías y versiones empleadas. El objetivo es proporcionar una visión exhaustiva de las herramientas y tecnologías que han permitido llevar a cabo el análisis y la modelización para la detección de cáncer a partir de datos de biomarcadores. Esta información es crucial para garantizar la reproducibilidad del estudio y ofrecer una guía clara sobre la infraestructura necesaria para replicar y ampliar los resultados obtenidos.

Arquitectura de Referencia

La arquitectura tecnológica del proyecto se compone de varias capas y componentes que interactúan entre sí para procesar, analizar y modelar los datos de biomarcadores. A continuación, se detalla cada una de estas capas y sus componentes principales.

Librerías importadas

```
# P
# Carga de librerías
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from scipy.stats import skew, kurtosis
from scipy.stats import shapiro
from scipy.stats import normaltest

# Entrenar el modelo
from sklearn.model_selection import train_test_split

# Selección de las variables por tipo
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder, LabelEncoder
from sklearn.compose import make_column_selector
from sklearn.impute import SimpleImputer
from sklearn.tree import DecisionTreeRegressor
from sklearn.feature_selection import mutual_info_classif

# Modelos
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, roc_curve, auc, confusion_matrix, accuracy_score,
precision_score, recall_score, f1_score, adjusted_rand_score, r2_score, silhouette_score,
davies_bouldin_score, calinski_harabasz_score
from sklearn.model_selection import cross_val_score
import matplotlib.pyplot as plt
```

1. Adquisición y Preprocesamiento de Datos

La primera capa se encarga de la adquisición y preprocesamiento de los datos. Los datos se obtuvieron de archivos en formato Excel y se preprocesaron para asegurar su calidad y coherencia.

- Componentes:
 - **Fuente de Datos:** Archivos Excel proporcionados ([Tables_S1_to_S11.xlsx](#)).
 - Librerías Utilizadas:
 - [pandas](#) (versión 1.3.3): Utilizada para la manipulación y análisis de datos.
 - [numpy](#) (versión 1.21.2): Utilizada para operaciones numéricas y manejo de matrices.
 - [openpyxl](#) (versión 3.0.7): Utilizada para la lectura de archivos Excel.

2. Análisis Exploratorio de Datos (EDA)

La segunda capa se enfoca en el análisis exploratorio de datos para entender mejor la distribución y las características de los datos de biomarcadores.

- Componentes:
 - Librerías Utilizadas:
 - [matplotlib](#) (versión 3.4.3): Utilizada para la visualización de datos.
 - [seaborn](#) (versión 0.11.2): Utilizada para la visualización de datos y gráficos estadísticos.

3. Modelización y Evaluación

La tercera capa abarca la modelización predictiva y la evaluación de los modelos utilizando diversas técnicas de aprendizaje automático.

- Componentes:
 - Librerías Utilizadas:
 - [scikit-learn](#) (versión 0.24.2): Utilizada para técnicas de modelización y evaluación.
 - [xgboost](#) (versión 1.4.2): Utilizada para la modelización avanzada con árboles de decisión.
 - [imbalanced-learn](#) (versión 0.8.0): Utilizada para el manejo de datos desbalanceados.

4. Infraestructura y Entorno de Desarrollo

El proyecto se ha desarrollado en un entorno basado en Jupyter Notebooks, facilitando la integración de código, visualizaciones y documentación.

- Componentes:
 - **Entorno de Desarrollo:** Jupyter Notebooks.
 - Librerías de Soporte:
 - [jupyter](#) (versión 1.0.0): Utilizada para la creación y ejecución de notebooks interactivos.

Conclusión

La arquitectura tecnológica utilizada en este proyecto abarca desde la adquisición y preprocesamiento de datos hasta la modelización y evaluación de resultados, pasando por un análisis exploratorio exhaustivo. Las diversas técnicas y herramientas empleadas han permitido una comprensión profunda de los datos de biomarcadores y la identificación de patrones relevantes para la detección de cáncer.

El uso de librerías robustas y entornos interactivos como Jupyter Notebooks ha facilitado el desarrollo y la replicabilidad del proyecto, asegurando que los resultados sean precisos y reproducibles. Este enfoque holístico proporciona una base sólida para futuras investigaciones y aplicaciones en el ámbito de la detección temprana del cáncer mediante análisis de biomarcadores sanguíneos.

Modelización

El objetivo de esta sección es detallar de manera exhaustiva y científica el proceso de modelización llevado a cabo para la detección de cáncer utilizando datos de biomarcadores. A continuación, se describen las técnicas utilizadas, el proceso de evaluación y los modelos construidos, junto con los resultados analíticos obtenidos.

Técnicas Utilizadas de Aprendizaje Supervisado

Se emplearon diversas técnicas de aprendizaje supervisado, abarcando desde modelos lineales hasta métodos de ensamblado y redes neuronales. Las principales técnicas utilizadas fueron:

1. Regresión Lineal
2. Regresión Logística
3. Árbol de Decisión
4. Bosques Aleatorios (Random Forest)

- 5. **KNN (K-Nearest Neighbors)**
- 6. **Máquinas de Soporte Vectorial (SVM)**
- 7. **Naive Bayes (Gaussian y Bernoulli)**
- 8. **AdaBoost**
- 9. **Gradient Boosting**
- 10. **Redes Neuronales (ANN, MLP, RNN)**
- 11. **Extreme Learning Machine (ELM)**
- 12. **Regresión Polinomial**

Cada técnica se evaluó en términos de precisión, recall, F1-score y otras métricas relevantes para asegurar la robustez y generalización de los modelos.

Proceso de Evaluación

Validación Cruzada

Para evaluar la validez de los modelos, se utilizó la validación cruzada, una técnica que permite medir el rendimiento del modelo de manera más precisa y confiable. Esta técnica divide el conjunto de datos en varios subconjuntos (folds) y entrena el modelo múltiples veces, cada vez utilizando un subconjunto diferente como conjunto de prueba y los restantes como conjunto de entrenamiento.

Las métricas de validación cruzada empleadas incluyeron:

- **Accuracy:** Proporción de predicciones correctas sobre el total de predicciones.
- **Precision:** Proporción de verdaderos positivos sobre el total de predicciones positivas.
- **Recall:** Proporción de verdaderos positivos sobre el total de positivos reales.
- **F1-Score:** Media armónica de precision y recall, proporcionando un balance entre ambas métricas.
- **AUC-ROC:** Área bajo la curva ROC, que evalúa la capacidad del modelo para distinguir entre clases.

Métricas de Evaluación

Además de la validación cruzada, se realizaron evaluaciones adicionales en conjuntos de datos separados para entrenamiento, validación y prueba. Las métricas utilizadas fueron:

- **MSE (Mean Squared Error)**
- **R-squared (R^2)**
- **Adjusted Rand Index**

Modelos

1. Regresión Lineal

- **Descripción:** La regresión lineal se utilizó para modelar la relación entre los biomarcadores y la probabilidad de detección de cáncer.
- **Resultados:**
 - **Training Set:** Accuracy: 79.52%, F1-Score: 79.50%
 - **Validation Set:** Accuracy: 81.59%, F1-Score: 81.55%
 - **Test Set:** Accuracy: 79.40%, F1-Score: 79.17%

2. Regresión Logística

- **Descripción:** Este modelo se empleó para predecir la presencia de cáncer mediante la probabilidad de un evento binario.
- **Resultados:**
 - **Training Set:** Accuracy: 83.56%, F1-Score: 83.63%
 - **Validation Set:** Accuracy: 82.69%, F1-Score: 82.71%
 - **Test Set:** Accuracy: 82.69%, F1-Score: 82.70%

3. Árbol de Decisión

- **Descripción:** Se optimizaron los hiperparámetros del árbol de decisión mediante búsqueda en rejilla.
- **Resultados:**
 - **Training Set:** Accuracy: 94.31%, F1-Score: 94.31%
 - **Validation Set:** Accuracy: 85.44%, F1-Score: 85.38%

- **Test Set:** Accuracy: 85.44%, F1-Score: 85.44%

4. Bosques Aleatorios (Random Forest)

- **Descripción:** Se utilizó RandomizedSearchCV para encontrar los mejores hiperparámetros.
- **Resultados:**
 - **Training Set:** Accuracy: 100.0%, F1-Score: 100.0%
 - **Validation Set:** Accuracy: 89.29%, F1-Score: 89.23%
 - **Test Set:** Accuracy: 90.93%, F1-Score: 90.91%

5. KNN (K-Nearest Neighbors)

- **Descripción:** Se evaluaron diferentes valores de K y se optimizaron los hiperparámetros.
- **Resultados:**
 - **Training Set:** Accuracy: 85.22%, F1-Score: 85.27%
 - **Validation Set:** Accuracy: 78.57%, F1-Score: 78.59%
 - **Test Set:** Accuracy: 79.40%, F1-Score: 79.43%

6. Máquinas de Soporte Vectorial (SVM)

- **Descripción:** Se utilizó GridSearchCV para encontrar los mejores parámetros.
- **Resultados:**
 - **Training Set:** Accuracy: 84.66%, F1-Score: 84.72%
 - **Validation Set:** Accuracy: 82.14%, F1-Score: 82.13%
 - **Test Set:** Accuracy: 82.14%, F1-Score: 82.18%

7. Naive Bayes (Gaussian y Bernoulli)

- **Descripción:** Se emplearon variantes de Gaussian y Bernoulli para evaluar su rendimiento en la detección de cáncer.
- **Resultados Gaussian:**
 - **Training Set:** Accuracy: 75.94%, F1-Score: 75.43%
 - **Validation Set:** Accuracy: 76.65%, F1-Score: 76.15%
 - **Test Set:** Accuracy: 75.82%, F1-Score: 75.24%
- **Resultados Bernoulli:**
 - **Training Set:** Accuracy: 81.45%, F1-Score: 81.52%
 - **Validation Set:** Accuracy: 79.12%, F1-Score: 79.11%
 - **Test Set:** Accuracy: 82.69%, F1-Score: 82.72%

8. AdaBoost

- **Descripción:** Se empleó un clasificador débil (Decision Tree) y se optimizaron los hiperparámetros.
- **Resultados:**
 - **Training Set:** Accuracy: 100.0%, F1-Score: 100.0%
 - **Validation Set:** Accuracy: 89.01%, F1-Score: 88.99%
 - **Test Set:** Accuracy: 92.86%, F1-Score: 92.86%

9. Gradient Boosting

- **Descripción:** Se optimizaron los hiperparámetros mediante búsqueda en rejilla.
- **Resultados:**
 - **Training Set:** Accuracy: 100.0%, F1-Score: 100.0%
 - **Validation Set:** Accuracy: 94.51%, F1-Score: 94.50%
 - **Test Set:** Accuracy: 93.13%, F1-Score: 93.13%

10. Redes Neuronales (ANN, MLP, RNN)

- **Descripción:** Se entrenaron diferentes arquitecturas de redes neuronales para capturar patrones complejos en los datos.
- **Resultados ANN:**
 - **Training Set:** Accuracy: 83.93%, F1-Score: 83.99%
 - **Validation Set:** Accuracy: 82.97%, F1-Score: 82.98%

- **Test Set:** Accuracy: 83.52%, F1-Score: 83.54%
- **Resultados MLP:**
 - **Training Set:** Accuracy: 84.22%, F1-Score: 84.45%
 - **Validation Set:** Accuracy: 83.24%, F1-Score: 83.26%
 - **Test Set:** Accuracy: 83.52%, F1-Score: 83.54%
- **Resultados RNN:**
 - **Training Set:** Accuracy: 85.31%, F1-Score: 85.33%
 - **Validation Set:** Accuracy: 81.87%, F1-Score: 81.81%
 - **Test Set:** Accuracy: 83.52%, F1-Score: 83.55%

11. Extreme Learning Machine (ELM)

- **Descripción:** Se utilizó ELM para entrenar un modelo de red neuronal con una única capa oculta.
- **Resultados:**
 - **Training Set:** Accuracy: 84.21%, F1-Score: 84.25%
 - **Validation Set:** Accuracy: 81.59%, F1-Score: 81.59

12. Perceptrón Multicapa (MLP)

- **Descripción:** El perceptrón multicapa (MLP) es un tipo de red neuronal feedforward que se entrenó para capturar patrones complejos en los datos de biomarcadores mediante varias capas ocultas con funciones de activación no lineales.
- **Resultados:**
 - **Training Set:** Accuracy: 84.39%, F1-Score: 84.45%
 - **Validation Set:** Accuracy: 83.24%, F1-Score: 83.26%
 - **Test Set:** Accuracy: 83.52%, F1-Score: 83.54%

13. Red Neuronal Recurrente (RNN)

- **Descripción:** Las redes neuronales recurrentes (RNN) fueron utilizadas para capturar dependencias temporales en los datos de biomarcadores. Se emplearon capas recurrentes con activación 'relu' para procesar las secuencias de datos.
- **Resultados:**
 - **Training Set:** Accuracy: 85.31%, F1-Score: 85.33%
 - **Validation Set:** Accuracy: 81.87%, F1-Score: 81.81%
 - **Test Set:** Accuracy: 83.52%, F1-Score: 83.55%
 - Estas técnicas avanzadas de redes neuronales demostraron ser efectivas en la modelización de datos complejos, proporcionando resultados sólidos y destacando su capacidad para capturar relaciones no lineales y temporales en los datos de biomarcadores para la detección de cáncer.

Técnicas Utilizadas de Aprendizaje No Supervisado

Se emplearon diversas técnicas de aprendizaje no supervisado, abarcando métodos de clustering y reducción de dimensionalidad. Las principales técnicas utilizadas fueron:

1. **KMeans**
2. **Mean Shift**
3. **DBSCAN**
4. **Gaussian Mixture Model (GMM)**
5. **Clustering Jerárquico**
6. **Algoritmo AnDE (Anomaly Detection Ensemble)**
7. **Detección de Anomalías (Isolation Forest)**
8. **Reducción de Dimensionalidad mediante SVD (Singular Value Decomposition)**
9. **Reducción de Dimensionalidad mediante PCA (Principal Component Analysis)**
10. **Análisis de Componentes Independientes (ICA)**
11. **Análisis Discriminante Lineal (LDA)**

Proceso de Evaluación

Validación Cruzada

Para evaluar la validez de los modelos no supervisados, se utilizó una combinación de métricas que permiten medir el rendimiento de los algoritmos de clustering y detección de anomalías de manera precisa y confiable. Las métricas de evaluación empleadas incluyeron:

- **Silhouette Score:** Mide cuán similares son los objetos en un clúster en comparación con los objetos de otros clústeres. Un valor más alto indica clústeres más definidos.
- **Davies-Bouldin Index:** Promedio de las tasas de similitud de cada clúster con el clúster más similar. Un valor más bajo indica clústeres más definidos.
- **Calinski-Harabasz Index:** Proporción de la suma de la dispersión entre clústeres y la dispersión dentro de los clústeres. Un valor más alto indica clústeres más definidos.

Métricas de Evaluación

Se realizaron evaluaciones adicionales en conjuntos de datos separados para entrenamiento, validación y prueba, utilizando las siguientes métricas:

- **Silhouette Score**
- **Davies-Bouldin Index**
- **Calinski-Harabasz Index**
- **Global Score:** Métrica compuesta que combina las anteriores para ofrecer una visión general del rendimiento del modelo.

Modelos

1. KMeans

- **Descripción:** KMeans es una técnica de clustering que agrupa los datos en K clústeres basados en la similitud de características.
- **Proceso:**
 - Se utilizó la técnica del codo y la puntuación de silueta para determinar el número óptimo de clústeres.
 - Se entrenó el modelo final con el número óptimo de clústeres utilizando los datos de entrenamiento.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.4276, Davies-Bouldin Index: 1.0653, Calinski-Harabasz Index: 215.2480
 - **Validation Set:** Silhouette Score: 0.5128, Davies-Bouldin Index: 1.1492, Calinski-Harabasz Index: 72.7247
 - **Test Set:** Silhouette Score: 0.4927, Davies-Bouldin Index: 0.9756, Calinski-Harabasz Index: 140.3259

2. Mean Shift

- **Descripción:** Mean Shift es una técnica de clustering que no requiere especificar el número de clústeres por adelantado. Se basa en encontrar los modos en la densidad de los datos.
- **Proceso:**
 - Se utilizó la búsqueda de ancho de banda óptimo en el conjunto de validación.
 - Se entrenó el modelo final con el ancho de banda óptimo en los datos de entrenamiento.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.4582, Davies-Bouldin Index: 0.5735, Calinski-Harabasz Index: 114.9748
 - **Validation Set:** Silhouette Score: 0.4821, Davies-Bouldin Index: 0.6987, Calinski-Harabasz Index: 74.6950
 - **Test Set:** Silhouette Score: 0.4413, Davies-Bouldin Index: 0.8014, Calinski-Harabasz Index: 68.3151

3. DBSCAN

- **Descripción:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es una técnica de clustering que encuentra clústeres de alta densidad y marca puntos como ruido si están en regiones de baja densidad.
- **Proceso:**
 - Se buscó el valor óptimo de epsilon utilizando el conjunto de entrenamiento.
 - Se entrenó el modelo final con el valor óptimo de epsilon.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.8349, Davies-Bouldin Index: 1.7217, Calinski-Harabasz Index: 170.5236
 - **Validation Set:** Silhouette Score: 0.8204, Davies-Bouldin Index: 1.6787, Calinski-Harabasz Index: 52.0399
 - **Test Set:** Silhouette Score: 0.8233, Davies-Bouldin Index: 1.1945, Calinski-Harabasz Index: 88.5002

4. Gaussian Mixture Model (GMM)

- **Descripción:** GMM es una técnica probabilística para modelar la distribución de los datos como una mezcla de múltiples distribuciones gaussianas.
- **Proceso:**
 - Se utilizó GridSearchCV para encontrar los mejores parámetros del modelo.
 - Se entrenó el modelo con los mejores parámetros encontrados.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.5892, Davies-Bouldin Index: 1.6353, Calinski-Harabasz Index: 215.7722
 - **Validation Set:** Silhouette Score: 0.6177, Davies-Bouldin Index: 1.7655, Calinski-Harabasz Index: 65.4537
 - **Test Set:** Silhouette Score: 0.6531, Davies-Bouldin Index: 1.4111, Calinski-Harabasz Index: 112.5643

5. PCA (Principal Component Analysis)

- **Descripción:** PCA es una técnica de reducción de dimensionalidad que transforma los datos a un nuevo espacio con menos dimensiones, preservando la mayor cantidad posible de varianza original.
- **Proceso:**
 - Se determinó el número óptimo de componentes principales que explican al menos el 95% de la varianza.
 - Se aplicó el modelo PCA a los datos y se utilizaron los datos transformados para el clustering con KMeans.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.5894, Davies-Bouldin Index: 1.6258, Calinski-Harabasz Index: 215.8816
 - **Validation Set:** Silhouette Score: 0.6225, Davies-Bouldin Index: 1.7704, Calinski-Harabasz Index: 65.4857
 - **Test Set:** Silhouette Score: 0.6531, Davies-Bouldin Index: 1.4111, Calinski-Harabasz Index: 112.5643

6. Análisis de Componentes Independientes (ICA)

- **Descripción:** ICA es una técnica de separación de señales que busca descomponer una señal multivariada en componentes estadísticamente independientes.
- **Proceso:**
 - Se aplicó ICA a los datos y se utilizó Isolation Forest para detectar anomalías en los datos transformados.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.6568, Davies-Bouldin Index: 2.0301, Calinski-Harabasz Index: 168.6725
 - **Validation Set:** Silhouette Score: 0.6507, Davies-Bouldin Index: 2.0672, Calinski-Harabasz Index: 56.1600
 - **Test Set:** Silhouette Score: 0.6903, Davies-Bouldin Index: 1.6556, Calinski-Harabasz Index: 93.0848

7. Clustering Jerárquico

- **Descripción:** El clustering jerárquico es una técnica que construye una jerarquía de clústeres mediante la fusión o división iterativa de clústeres.
- **Proceso:**
 - Se determinó el número óptimo de clústeres utilizando el conjunto de entrenamiento.
 - Se entrenó el modelo final con el número óptimo de clústeres.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.8363, Davies-Bouldin Index: 1.3482, Calinski-Harabasz Index: 197.3254
 - **Validation Set:** Silhouette Score: 0.8887, Davies-Bouldin Index: 0.0743, Calinski-Harabasz Index: 82.0443
 - **Test Set:** Silhouette Score: 0.6206, Davies-Bouldin Index: 1.4987, Calinski-Harabasz Index: 100.8270

8. Algoritmo AnDE

- **Descripción:** AnDE (Anomaly Detection Ensemble) es una técnica para detectar anomalías basada en la densidad de los puntos y un umbral determinado.
- **Proceso:**
 - Se entrenó el modelo AnDE con los datos de entrenamiento y se calculó el umbral óptimo basado en la densidad.
 - Se aplicó el modelo para detectar anomalías en los conjuntos de entrenamiento, validación y prueba.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.7439, Davies-Bouldin Index: 2.0779, Calinski-Harabasz Index: 157.0921
 - **Validation Set:** Silhouette Score: 0.6827, Davies-Bouldin Index: 2.0878, Calinski-Harabasz Index: 53.4775
 - **Test Set:** Silhouette Score: 0.7369, Davies-Bouldin Index: 1.6303, Calinski-Harabasz Index: 91.8491

9. Detección de Anomalías (Isolation Forest)

- **Descripción:** Isolation Forest es una técnica de detección de anomalías que construye árboles de aislamiento para identificar puntos de datos anómalos.
- **Proceso:**
 - Se entrenó el modelo Isolation Forest con los datos de entrenamiento.
 - Se detectaron anomalías en los conjuntos de entrenamiento, validación y prueba.
- **Resultados:**
 - **Training Set:** Silhouette Score: 0.6577, Davies-Bouldin Index: 1.9636, Calinski-Harabasz Index: 177.3739
 - **Validation Set:** Silhouette Score: 0.6336, Davies-Bouldin Index: 2.0642, Calinski-Harabasz Index

10. Reducción de Dimensionalidad SVD

- **Descripción:** La técnica de Reducción de Dimensionalidad mediante Descomposición en Valores Singulares (SVD) se utiliza para reducir el número de características en los datos, manteniendo la mayor cantidad de información posible.
- **Proceso:**
 1. **Determinación de Componentes Óptimos:**
 - Se ajustó un modelo SVD preliminar para calcular la varianza explicada acumulada.
 - Se seleccionó el número de componentes que explican al menos el 95% de la varianza.
 2. **Aplicación del Modelo SVD:**
 - Se transformaron los datos de entrenamiento, validación y prueba al nuevo espacio reducido utilizando el modelo SVD ajustado.
 3. **Clustering con KMeans:**
 - Se determinó el número óptimo de clústeres utilizando la técnica del codo y la puntuación de silueta.
 - Se aplicó el algoritmo KMeans a los datos reducidos para realizar el clustering.
- **Resultados:**
 - **Training Set:**
 - Silhouette Score: 0.5894
 - Davies-Bouldin Index: 1.6258
 - Calinski-Harabasz Index: 215.8816
 - **Validation Set:**
 - Silhouette Score: 0.6225
 - Davies-Bouldin Index: 1.7704
 - Calinski-Harabasz Index: 65.4857
 - **Test Set:**
 - Silhouette Score: 0.6531
 - Davies-Bouldin Index: 1.4111
 - Calinski-Harabasz Index: 112.5643

Bibliografía y recursos

1. Wong, K.-C., Chen, J., Zhang, J., Lin, J., Yan, S., Zhang, S., Li, X., Liang, C., Peng, C., Lin, Q., Kwong, S., & Yu, J. (2019). Early Cancer Detection from Multianalyte Blood Test Results. *iScience*, 15, 332-341. <https://doi.org/10.1016/j.isci.2019.04.035>
2. Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A. A., Wong, F., Mattox, A., Hruban, R. H., Wolfgang, C. L., Goggins, M. G., Dal Molin, M., Wang, H., Roden, R., Eshleman, J. R., Husain, H., Lennon, A. M., ... & Vogelstein, B. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378), 926-930. <https://doi.org/10.1126/science.aar3247>
3. Smith, R. A., Andrews, K. S., Brooks, D., Fedewa, S. A., Manassaram-Baptiste, D., Saslow, D., & Wender, R. C. (2019). Cancer screening in the United States, 2019: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA: A Cancer Journal for Clinicians*, 69(3), 184-210. <https://doi.org/10.3322/caac.21557>
4. Hackshaw, A., Clarke, C. A., & Hartman, A. R. (2022). New genomic technologies for multi-cancer early detection: Rethinking the scope of cancer screening. *Cancer Cell*, 40(2), 109-113. <https://doi.org/10.1016/j.ccell.2022.01.005>

5. LeeVan, E., & Pinsky, P. (2024). Predictive performance of cell-free nucleic acid-based multi-cancer early detection tests: a systematic review. *Clinical Chemistry*, 70(1), 90-101. <https://doi.org/10.1093/clinchem/hvaa190>
6. Klein, E. A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., ... & Curtis, C. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology*, 32(9), 1167-1177. <https://doi.org/10.1016/j.annonc.2021.06.003>
7. Multi-cancer early detection tests: pioneering a revolution in cancer screening. *Clinical Cancer Bulletin*. <https://link.springer.com/article/10.1007/s10555-022-09965-4>
8. Multi-cancer Early Detection Blood Tests (MCED) Debut - DNA Science. *PLoS*. <https://dnascience.plos.org/article/doi/10.1371/journal.pone.0274855>
9. Smith, R. A., Andrews, K. S., Brooks, D., et al. (2019). Cancer screening in the United States, 2019: a review of current American Cancer Society guidelines and current issues in cancer screening. *CA Cancer J Clin*, 69(3), 184-210. <https://doi.org/10.3322/caac.21557>
10. Hackshaw, A., Clarke, C. A., Hartman, A. R. (2022). New genomic technologies for multi-cancer early detection: Rethinking the scope of cancer screening. *Cancer Cell*, 40(2), 109-113. <https://doi.org/10.1016/j.ccell.2022.01.005>
11. Cohen, J. D., Li, L., Wang, Y., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 359(6378), 926-930. <https://doi.org/10.1126/science.aar3247>
12. LeeVan, E., Pinsky, P. (2024). Predictive performance of cell-free nucleic acid-based multi-cancer early detection tests: a systematic review. *Clinical Chemistry*, 70(1), 90-101. <https://doi.org/10.1093/clinchem/hvaa190>
13. Klein, E. A., Richards, D., Cohn, A., et al. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology*, 32(9), 1167-1177. <https://doi.org/10.1016/j.annonc.2021.06.003>
14. Early Cancer Detection from Multianalyte Blood Test Results. *PMC*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548890/>