KSCHOOL

Trabajo Fin de Máster

Máster Data Science

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

Autores:

Florina Cretu
Mar Benitez de Lucio Villegas
Kevin Jhoan Orozco Agudelo
Daniel del Río Alonso

Tutor:

Antonio Pita Lozano

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

Abstract

La detección temprana del cáncer incrementa significativamente las probabilidades de un tratamiento exitoso y la supervivencia del paciente. Este proyecto tiene como finalidad replicar y mejorar los hallazgos del estudio "Early Cancer Detection from Multianalyte Blood Test Results", utilizando análisis de sangre multianalito para predecir tanto la presencia como el tipo de cáncer. El estudio emplea diversos modelos de aprendizaje automático para analizar marcadores sanguíneos y aumentar la precisión en la detección y clasificación del cáncer.

El proyecto persigue dos objetivos principales. En primer lugar, buscamos replicar los modelos del estudio original para predecir la presencia de cáncer en pacientes basándonos en los resultados de análisis de sangre. En segundo lugar, pretendemos extender estos modelos para no solo detectar la presencia de cáncer, sino también para clasificar el tipo específico de cáncer entre ocho categorías predefinidas. Nuestro enfoque incluye el desarrollo y validación de nuevos modelos con el fin de mejorar el rendimiento predictivo más allá de los resultados iniciales.

Para alcanzar estos objetivos, utilizamos datos de dos conjuntos clave: uno destinado a la detección binaria del cáncer (cáncer vs. no cáncer) y otro enfocado en la clasificación multicategoría del cáncer (identificación de tipos específicos de cáncer). Empleamos una amplia gama de algoritmos de aprendizaje supervisado, incluyendo variantes de Naive Bayes, árboles de decisión y modelos de aprendizaje profundo. Prestamos 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, manteniendo 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 aporta significativamente 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 puedan 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

Introducción

Contexto y Motivación

La detección temprana del cáncer es fundamental 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, los cuales pueden ser costosos y poco accesibles para toda la población. Sin embargo, 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. Esto permite una detección menos invasiva y potencialmente más temprana del cáncer (DNA Science) (AACR Journals).

Estado del Arte

En los últimos años, se han logrado avances significativos en la detección temprana del cáncer mediante análisis de sangre multianalito. Estos progresos 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 pruebas de detección temprana de múltiples tipos de cáncer (MCEDs) ha mostrado un gran 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. Entre estos desafíos se 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).

Existen varios casos de uso potenciales para un sistema de detección temprana de cáncer basado en análisis de sangre multianalito. Entre ellos se incluyen:

1. **Detección Temprana en Poblaciones de Alto Riesgo:** Implementación de pruebas regulares en individuos con antecedentes familiares de cáncer o con otros factores de riesgo conocidos.
2. **Monitoreo Post-Tratamiento:** Seguimiento de pacientes que han recibido tratamiento para el cáncer para detectar posibles recurrencias de manera temprana.
3. **Cribado General:** Uso de este sistema como una herramienta de cribado en chequeos médicos rutinarios para la población general.

Retos Similares

La implementación de sistemas de diagnóstico basados en datos no está exenta de desafíos. Entre los principales retos se encuentran:

1. **Calidad y Consistencia de los Datos:** Garantizar que los datos de las pruebas sanguíneas sean de alta calidad y consistentes es fundamental para el éxito del modelo predictivo.
2. **Variabilidad Biológica:** Los niveles de biomarcadores pueden variar significativamente entre individuos debido a factores no relacionados con el cáncer, lo que puede complicar la detección precisa.
3. **Interpretabilidad del Modelo:** Los modelos de aprendizaje automático deben ser interpretables para ser aceptados en un entorno clínico, donde los profesionales de la salud necesitan comprender las razones detrás de una predicción.

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 en 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. Este proceso implica utilizar datos de múltiples marcadores sanguíneos y aplicar algoritmos avanzados de aprendizaje supervisado para mejorar la precisión diagnóstica. La clasificación del tipo de cáncer se realizará mediante la aplicación del teorema de la probabilidad total, lo que permitirá una evaluación más precisa y robusta de los diferentes tipos de cáncer.

El teorema de la probabilidad total es una regla fundamental en la teoría de la probabilidad que permite calcular la probabilidad total de un evento a partir de las probabilidades condicionadas de dicho evento en diferentes escenarios. Fue desarrollado por el matemático suizo Jakob Bernoulli en el siglo XVIII. Este teorema se utiliza para descomponer un problema complejo en partes más simples, lo que facilita el análisis y la solución de problemas en diversas disciplinas, incluyendo la detección y clasificación del cáncer en este proyecto.

Contribuciones Esperadas

Las contribuciones esperadas de este proyecto incluyen:

1. **Replicación y Validación de Modelos Existentes**: La validación de modelos de predicción existentes es crucial para confirmar la eficacia de los métodos desarrollados en estudios previos. Al replicar los resultados obtenidos en estudios como el presentado en el artículo "Multi-Cancer Early Detection Blood Tests (MCED)" publicado en DNA Science, este proyecto refuerza la robustez y reproducibilidad de los modelos utilizados. Este paso es esencial para asegurar que las metodologías propuestas pueden ser aplicadas consistentemente en diferentes entornos y con diferentes conjuntos de datos.
2. **Desarrollo de Nuevos Modelos**: Además de validar los modelos existentes, el proyecto se enfoca en el desarrollo y prueba de nuevos modelos de aprendizaje automático que puedan superar la precisión y la eficacia de los modelos actuales. La investigación en esta área podría identificar enfoques novedosos o combinaciones de técnicas que mejoren la capacidad predictiva y reduzcan la tasa de falsos positivos y negativos. Según un artículo reciente en SpringerLink, la innovación en algoritmos de detección temprana de cáncer puede ofrecer mejoras significativas en el rendimiento diagnóstico.
3. **Análisis Comparativo**: Realizar un análisis comparativo exhaustivo de diferentes enfoques y algoritmos es fundamental para determinar las mejores prácticas en la detección temprana del cáncer. Este análisis incluirá métodos supervisados y no supervisados, y considerará múltiples métricas de evaluación para ofrecer una visión completa de la eficacia de cada enfoque. Tal comparación no solo identificará los métodos más efectivos, sino que también destacará las condiciones en las que cada algoritmo presenta un mejor desempeño, facilitando su aplicación en diferentes escenarios clínicos.
4. **Aplicación Práctica**: Una de las metas principales del proyecto es proporcionar una base sólida para la implementación práctica de estas tecnologías en entornos clínicos. Esto incluye el desarrollo de protocolos y guías que permitan a los profesionales de la salud utilizar estas herramientas de manera efectiva para mejorar la detección temprana y el tratamiento del cáncer. La integración de estos avances en la práctica clínica podría resultar en diagnósticos más rápidos y precisos, optimizando el tratamiento y mejorando los resultados para los pacientes. Artículos como el de SpringerLink y AACR Journals subrayan la importancia de llevar estos desarrollos tecnológicos del laboratorio a la práctica clínica.
5. **Propuestas para Nuevos Protocolos Clínicos**: Los hallazgos de este proyecto podrían influir en la creación de nuevos protocolos clínicos para la detección temprana del cáncer. La implementación de pruebas regulares basadas en análisis de sangre multianalito podría convertirse en una práctica estándar en chequeos médicos, especialmente para poblaciones de alto riesgo. Esto no solo mejorará los resultados de los pacientes, sino que también optimizará los recursos del sistema de salud.
6. **Generación de Conocimiento para Futuras Investigaciones**: Los resultados y datos generados por este proyecto serán de gran valor para futuras investigaciones. Al proporcionar un conjunto de datos exhaustivo y bien documentado, junto con los modelos y técnicas desarrollados, se facilitará la replicación y la extensión de este trabajo por parte de otros investigadores. Esto promoverá una mayor colaboración y avance en el campo de la detección temprana del cáncer.
7. **Impacto Económico y Social**: La implementación exitosa de este sistema de detección temprana podría tener un impacto económico significativo al reducir los costos asociados con el tratamiento del cáncer avanzado. Además, desde una perspectiva social, mejorar las tasas de detección temprana contribuirá a una mejor calidad de vida para los pacientes y sus familias, al aumentar las tasas de supervivencia y reducir la carga emocional y física del tratamiento del cáncer.

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) (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).
4. **Inserción en el Mercado y Adopción Global:**
    - **Objetivo**: Introducir el sistema de detección temprana del cáncer en mercados globales clave, asegurando una adopción amplia en clínicas y hospitales.
    - **Justificación**: Ampliar la presencia del producto en mercados internacionales contribuirá a mejorar el acceso de los pacientes a pruebas de detección avanzadas.
5. **Partnerships Estratégicas y Colaboraciones**:
    - **Objetivo**: Establecer alianzas estratégicas con laboratorios de renombre y centros de investigación para validar y promover la tecnología de detección del cáncer.
    - **Justificación**: Las colaboraciones con expertos en investigación y desarrollo fortalecerán la credibilidad del producto y acelerarán su integración en prácticas clínicas.
6. **Cumplimiento Normativo y Regulatorio**:
    - **Objetivo**: Obtener las aprobaciones regulatorias necesarias (como la FDA en Estados Unidos o la CE en Europa) para comercializar la tecnología de detección del cáncer.
    - **Justificación**: Cumplir con los estándares regulatorios es crucial para asegurar la confianza del mercado y la aceptación del producto por parte de los profesionales de la salud.
7. **Educación y Concientización Pública**:
    - **Objetivo**: Educación pública y profesional sobre la importancia de la detección temprana del cáncer y las ventajas de las nuevas tecnologías de análisis de sangre.
    - **Justificación**: Aumentar la conciencia pública y profesional puede conducir a una mayor demanda y aceptación del producto, así como fomentar prácticas de salud preventiva.

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:**
    - **Descripción**: Lograr una sensibilidad y especificidad significativamente superiores a las de los métodos actuales de detección del cáncer.
    - **Justificación**: Una mejora en estos parámetros indica una mayor capacidad para identificar correctamente los casos positivos y negativos, reduciendo así los falsos diagnósticos y mejorando la confianza en las pruebas.
2. **Costos Reducidos de Pruebas de Diagnóstico:**
    - **Descripción**: Demostrar que el método desarrollado es más económico que las técnicas de diagnóstico tradicionales.
    - **Justificación**: La reducción de costos facilita una mayor adopción de la tecnología en diversas regiones, especialmente en áreas con recursos limitados, y puede hacer que las pruebas sean más accesibles para una mayor cantidad de personas.

3. **Adopción Clínica y Escalabilidad:**
    - **Descripción**: Probar que el sistema de detección es fácilmente implementable en entornos clínicos y puede ser escalado para su uso a gran escala.
    - **Justificación**: La facilidad de implementación y la capacidad de escalar el sistema son cruciales para su adopción en la práctica clínica diaria, asegurando que los avances tecnológicos se traduzcan en mejoras reales en la atención médica.

4. **Validación Cruzada Independiente:**
    - **Descripción**: Conseguir resultados positivos en estudios de validación cruzada independientes utilizando conjuntos de datos diversos.
    - **Justificación**: La validación cruzada independiente asegura la robustez y generalizabilidad de los modelos desarrollados, demostrando que pueden ser efectivos en diferentes contextos y poblaciones.

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. Este proyecto aspira a mejorar la calidad de vida de los pacientes mediante la innovación tecnológica y la implementación práctica de nuevos métodos de diagnóstico.

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: 31 variables.

3. **Tipología e Interpretación de las Variables:**
    - **Primer Conjunto de Datos**:
        - **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.
    - **Segundo Conjunto de Datos**:

      Para este conjunto de datos, además de las variables anteriores, se incluyen concentraciones de 31 marcadores proteicos adicionales, entre ellos:

        - **TGFa**: Factor de crecimiento transformante alfa, relevante en varios tipos de cáncer.
        - **HE4**: Proteína epidídimo humana 4, utilizada principalmente en el cáncer de ovario.
        - **sFas**: Receptor de muerte soluble, involucrado en la apoptosis celular.
        - **Thrombospondin-2**: Glicoproteína implicada en la regulación de la angiogénesis.
        - **AFP**: Alfa-fetoproteína, marcador para el cáncer de hígado.
        - **G-CSF**: Factor estimulante de colonias de granulocitos, utilizado en diversos tipos de cáncer.
        - **IL-6**: Interleucina 6, una citocina proinflamatoria.

      Entre otros marcadores relevantes para la clasificación de tipos específicos de cáncer.

NOTEBOOK 1: Modelos de Clasificación Binaria del Cáncer (Cáncer vs. no 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):

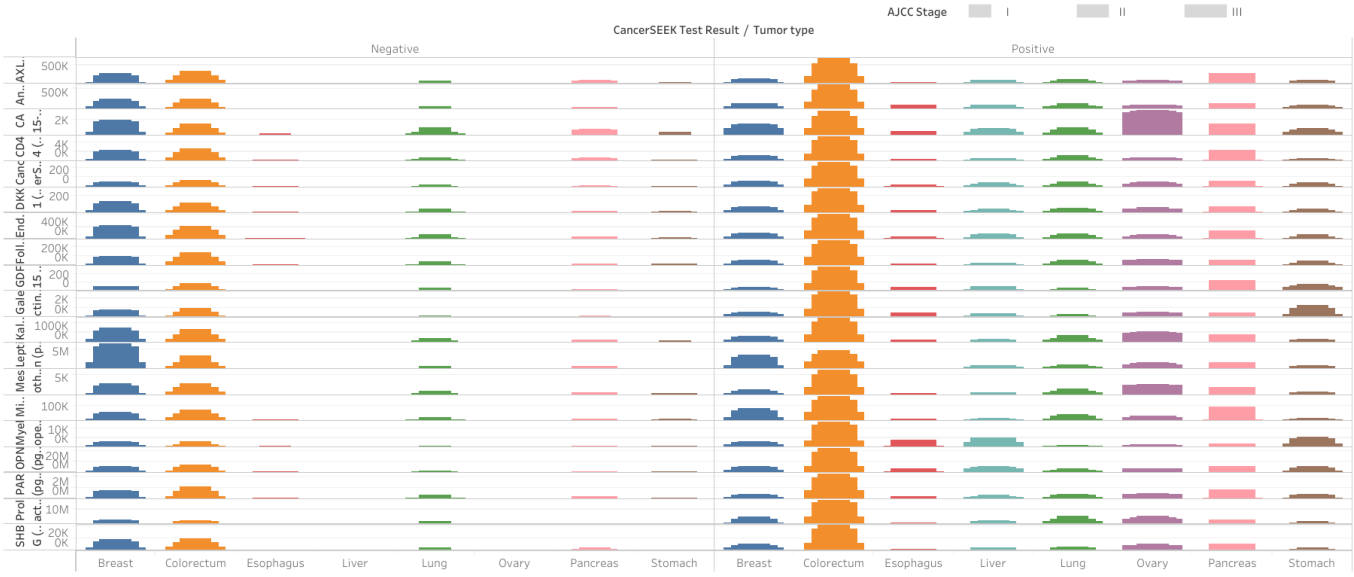
- **Descripción**: Mieloperoxidasa, enzima implicada en la inflamación y cáncer.
- **Media**: 31.20 ng/ml.
- **Desviación estándar**: 68.26 ng/ml.
- **Valores mínimos y máximos**: Rango de 1.3 a 1001 ng/ml.
- **Rango intercuartílico (IQR)**: 8.05 a 22.63 ng/ml.

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.

2. **Visualización de los Datos Sin Procesar**:



Datos sin procesar, ver en web para más información

Se realizó una visualización preliminar de los datos sin procesar para entender mejor su distribución y características. Este paso es crucial para identificar patrones, anomalías y distribuciones específicas de las variables, así como para planificar los pasos de preprocesamiento necesarios.

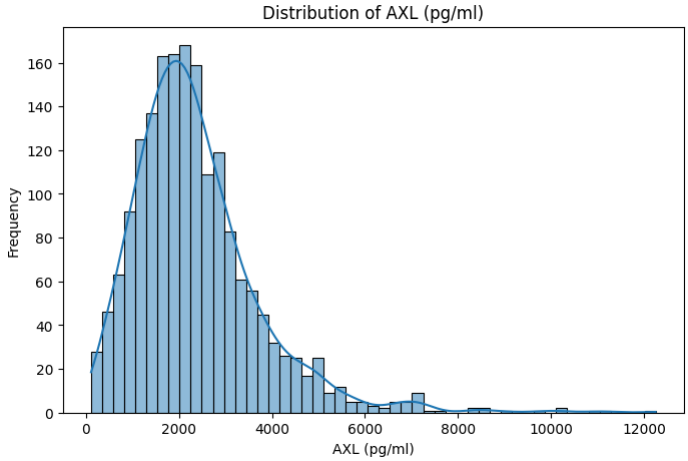
Análisis de Valores Nulos

Se observó que el parámetro "AJCC stage" presenta valores nulos en varias tablas. Esto no es sorprendente ya que este parámetro describe la severidad del cáncer y su ausencia en algunos registros puede indicar que los pacientes estaban libres de cáncer en el momento del estudio. Una situación similar se presenta con la variable "Histopatología" en el DataFrame 4. Para estas variables, podría considerarse la imputación de valores consensuados en lugar de dejarlas como nulas.

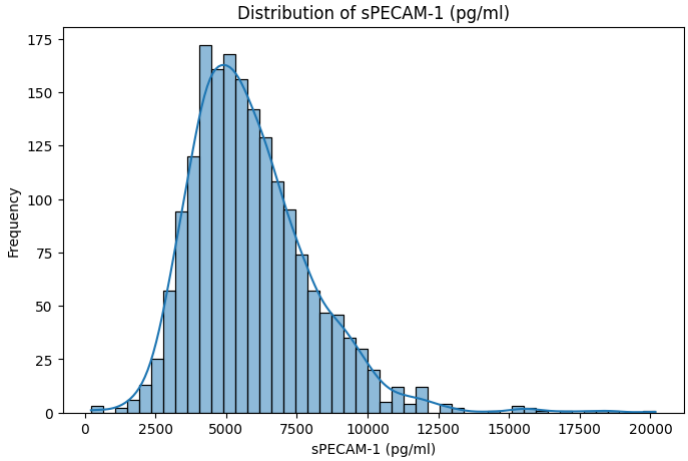
Distribución de Variables Numéricas

Se procedió a visualizar la distribución de varias variables numéricas clave del conjunto de datos. A continuación se describen los histogramas generados para cada una de estas variables:

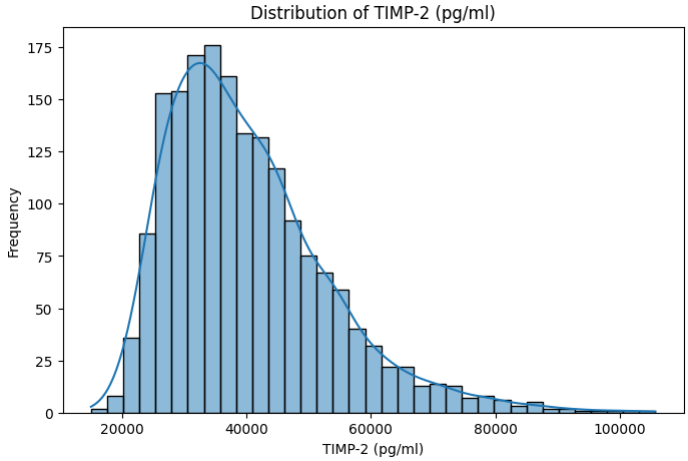
1. **Distribución de AXL (pg/ml)**:
    - **Descripción**: La distribución de la variable AXL muestra una forma sesgada hacia la derecha, lo cual es típico en datos biológicos donde una pequeña fracción de la población puede presentar valores extremadamente altos. La mayoría de los valores se concentran entre 0 y 5000 pg/ml, con una disminución gradual en frecuencia a medida que los valores aumentan.

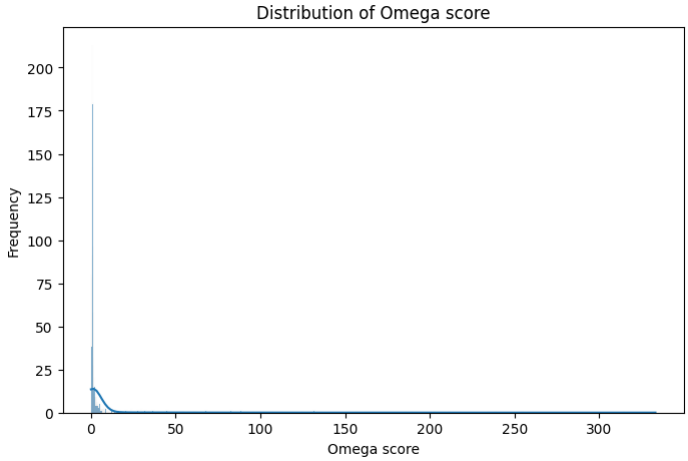


2. **Distribución de sPECAM-1 (pg/ml)**:
    - **Descripción**: La distribución de la variable sPECAM-1 tiene una forma similar a una distribución normal pero con cola larga a la derecha. La mayoría de los valores se encuentran entre 2500 y 12500 pg/ml, siendo la mediana alrededor de 5000 pg/ml.

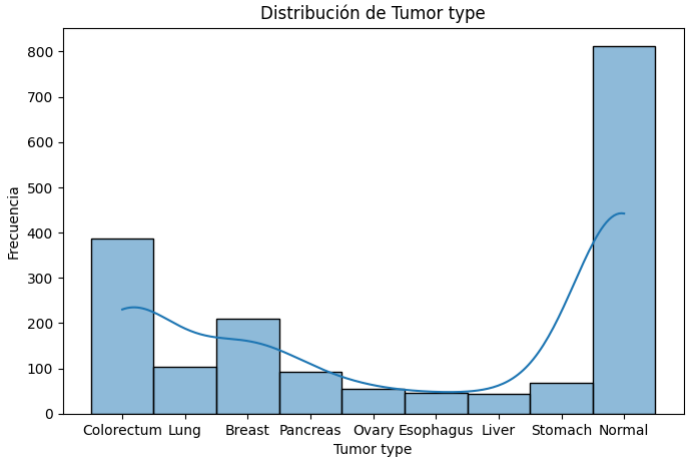


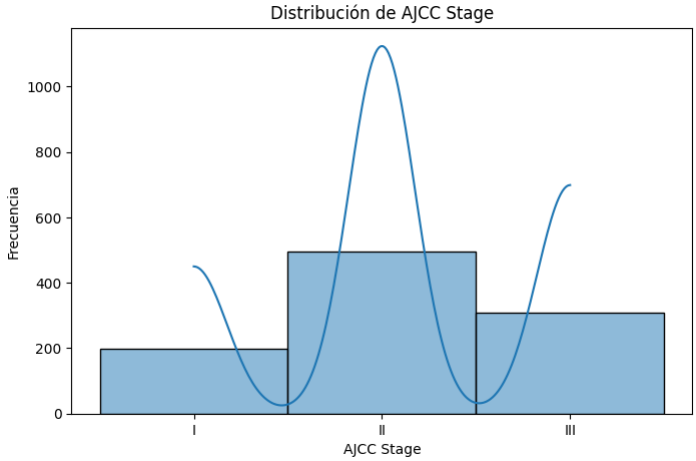
3. **Distribución de TIMP-2 (pg/ml)**:
    - **Descripción**: La variable TIMP-2 también muestra una distribución sesgada hacia la derecha. La mayor parte de los datos se agrupa entre 20000 y 80000 pg/ml, con una disminución notable en frecuencia para valores superiores.



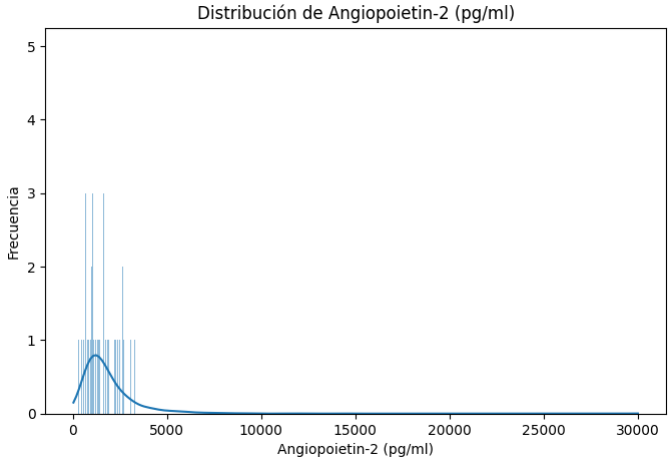
4. **Distribución de Omega Score**:
    - **Descripción**: El Omega Score presenta una distribución altamente sesgada hacia la derecha, con la mayoría de los valores muy cerca de 0 y pocos valores extremadamente altos. Este tipo de distribución puede indicar la presencia de valores atípicos o una escala logarítmica subyacente en los datos.
    - 

Distribución de las variables categóricas

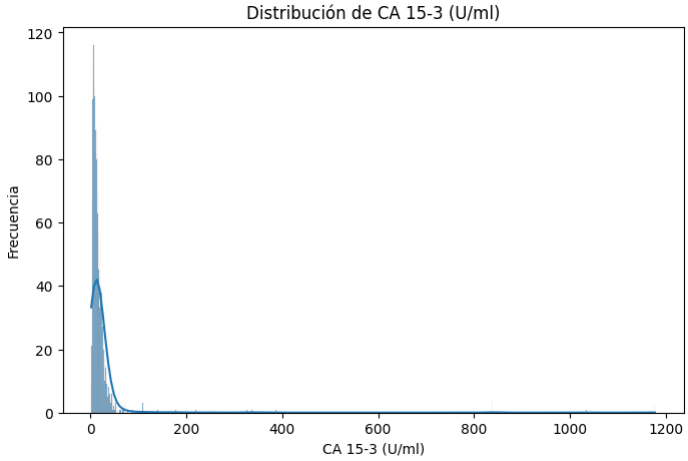
1. **Distribución de Tumor type**:
    - Este gráfico muestra la distribución de los diferentes tipos de tumores.
    - El tipo de tumor más frecuente es el "Normal" seguido por "Colorectum" y "Breast".
    - Otros tipos de tumores como "Lung", "Pancreas", "Ovary", "Esophagus", "Liver", y "Stomach" tienen una menor representación en el conjunto de datos.
    - 

    - Este gráfico muestra la distribución de las etapas AJCC.
    - La mayoría de los casos se encuentran en la etapa II, seguida por la etapa III y la etapa I.
    - Esto sugiere que la mayoría de los pacientes en el estudio fueron diagnosticados en etapas intermedias o avanzadas de cáncer.
    - 

2. **Distribución de Angiopoietin-2 (pg/ml)**:
    - Este gráfico muestra la distribución de los valores de Angiopoietin-2 en picogramos por mililitro.
    - La distribución es altamente asimétrica hacia la derecha, con la mayoría de los valores concentrados cerca de cero.
    - Hay pocos valores extremadamente altos, lo que indica la presencia de algunos outliers significativos.



    - Este gráfico muestra la distribución de los valores de CA 15-3 en unidades por mililitro.
    - La distribución es altamente asimétrica hacia la derecha, con la mayoría de los valores concentrados cerca de cero.
    - Hay pocos valores extremadamente altos, lo que indica la presencia de algunos outliers significativos.



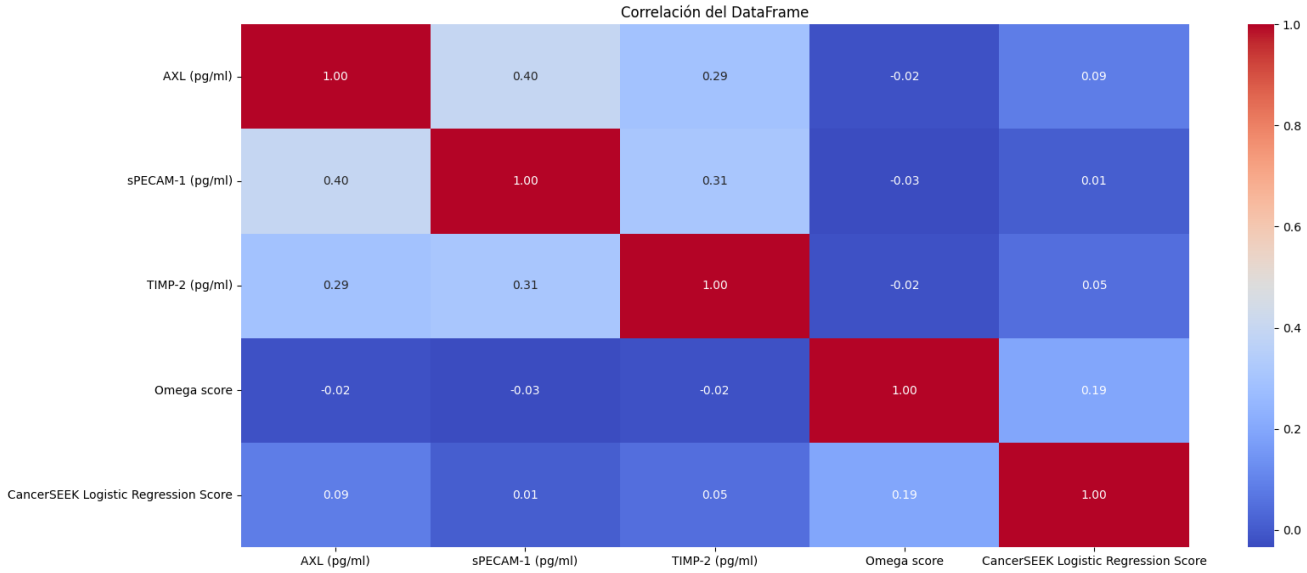
Correlaciones entre variables:

En el análisis de datos, la identificación de correlaciones entre variables es fundamental para entender cómo se relacionan entre sí y cómo pueden influir en los resultados de los modelos predictivos. La correlación mide la fuerza y la dirección de la relación lineal entre dos variables. Un coeficiente de correlación puede variar entre -1 y 1, donde:

- 1 indica una correlación positiva perfecta,
- -1 indica una correlación negativa perfecta, y
- 0 indica que no hay correlación lineal.

Para nuestro estudio sobre la predicción del cáncer, hemos analizado las correlaciones entre varias variables tanto numéricas como categóricas. A continuación, se presentan y explican los gráficos de correlación generados.

1. **Correlación entre Variables Numéricas**



En este gráfico de calor se muestran las correlaciones entre las siguientes variables numéricas:

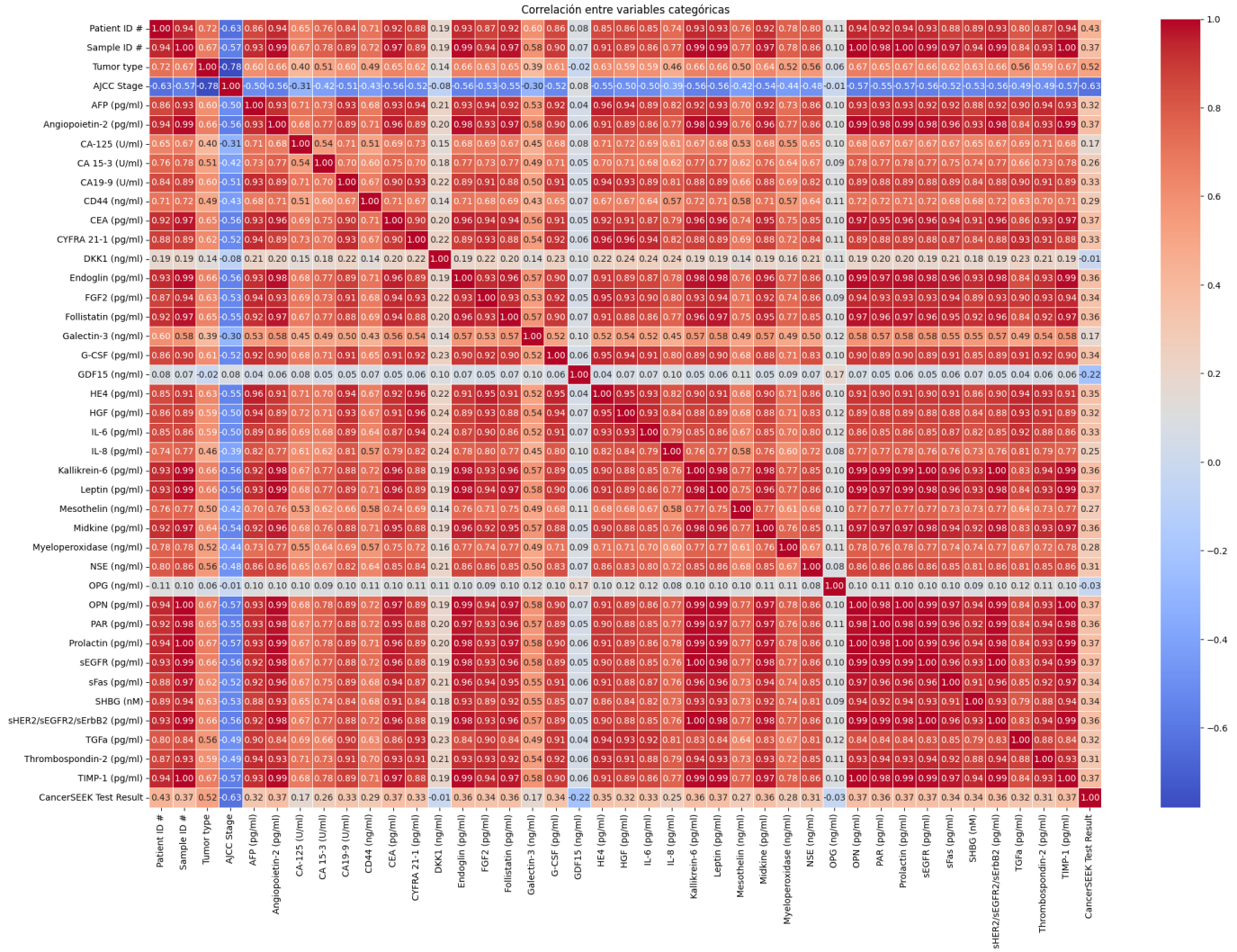
- AXL (pg/ml)
- sPECAM-1 (pg/ml)
- TIMP-2 (pg/ml)
- Omega score

Las principales observaciones de este gráfico son:

1. **AXL y sPECAM-1**: Hay una correlación positiva moderada (r ≈ 0.40), lo que indica que a medida que los niveles de AXL aumentan, los niveles de sPECAM-1 también tienden a aumentar.
2. **sPECAM-1 y TIMP-2**: También presentan una correlación positiva moderada (r ≈ 0.31).
3. **AXL y TIMP-2**: La correlación entre estos dos marcadores es más baja pero aún positiva (r ≈ 0.29).

En general, estas correlaciones indican algunas relaciones lineales positivas entre los biomarcadores estudiados y las puntuaciones de riesgo de cáncer.

2. **Correlación entre Variables Categóricas**



Este gráfico de calor muestra las correlaciones entre una amplia gama de variables categóricas y numéricas del estudio. Las variables incluidas abarcan desde identificadores de paciente y tipo de muestra hasta diversos marcadores tumorales y proteínas.

Las principales observaciones de este gráfico son:

1. **ID del Paciente y ID de Muestra**: La correlación es alta (r ≈ 0.50), lo que es esperado ya que cada muestra está vinculada a un paciente específico.
2. **Tumor type y varios marcadores**: Se observan varias correlaciones moderadas con distintos marcadores tumorales, lo que sugiere que ciertos tipos de tumores tienen perfiles biomarcadores característicos.

Matriz de Correlación Detallada

En la matriz de correlación detallada se identificaron las siguientes relaciones notables:

variable_1	variable_2	r	abs_r
AXL (pg/ml)	sPECAM-1 (pg/ml)	0.396347	0.396347
sPECAM-1 (pg/ml)	AXL (pg/ml)	0.396347	0.396347
sPECAM-1 (pg/ml)	TIMP-2 (pg/ml)	0.305282	0.305282
TIMP-2 (pg/ml)	sPECAM-1 (pg/ml)	0.305282	0.305282
TIMP-2 (pg/ml)	AXL (pg/ml)	0.289164	0.289164
AXL (pg/ml)	TIMP-2 (pg/ml)	0.289164	0.289164
sPECAM-1 (pg/ml)	Omega score	-0.034447	0.034447
Omega score	sPECAM-1 (pg/ml)	-0.034447	0.034447
Omega score	TIMP-2 (pg/ml)	-0.021214	0.021214
TIMP-2 (pg/ml)	Omega score	-0.021214	0.021214
Omega score	AXL (pg/ml)	-0.019419	0.019419
AXL (pg/ml)	Omega score	-0.019419	0.019419

Estos valores de correlación muestran cómo se relacionan los diferentes biomarcadores y las puntuaciones de los tests, proporcionando información valiosa para la modelización predictiva y la comprensión de las interacciones biológicas subyacentes.

Análisis profundo de los datos

Se realiza un análisis exhaustivo de las variables numéricas del conjunto de datos, verificando su distribución, asimetría y curtosis. Esto es fundamental para entender la naturaleza de los datos y para aplicar transformaciones adecuadas que mejoren el rendimiento de los modelos predictivos. A continuación, se explica el proceso y los resultados obtenidos.

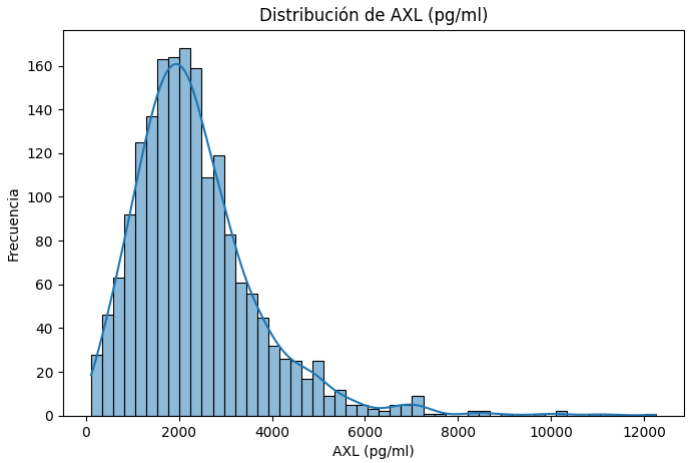
Visualización de la Distribución de las Variables Numéricas

Para cada variable numérica del conjunto de datos `df6`, se visualiza su distribución mediante histogramas con un ajuste de densidad de kernel (KDE). Adicionalmente, se calculan la asimetría y la curtosis de cada distribución, y se realizan pruebas de normalidad para determinar si las distribuciones siguen una distribución normal.

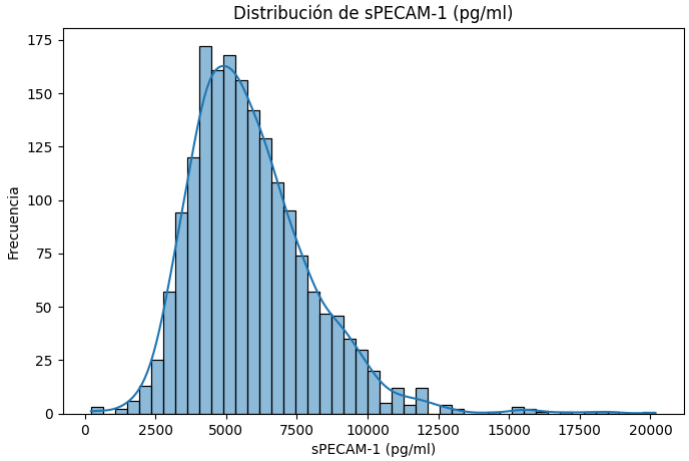
- Resultados de la Distribución

A continuación, se presentan y comentan los gráficos obtenidos para algunas de las variables más relevantes:

1. **AXL (pg/ml):**
    - Gráfico:

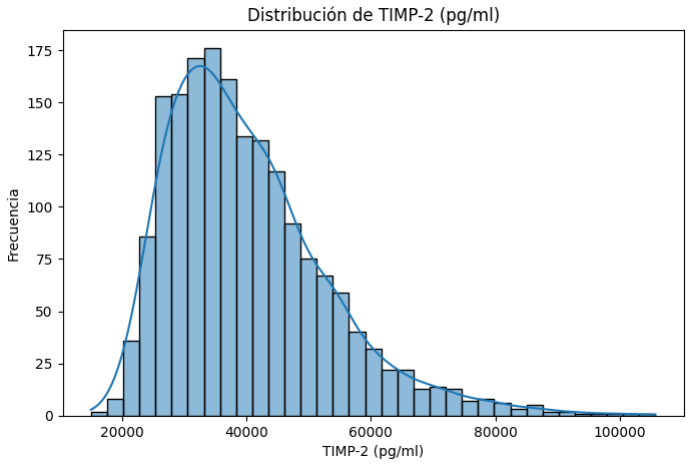


    - **Descripción:** La distribución de AXL muestra una asimetría positiva significativa, lo que indica que la mayoría de los valores se concentran en la parte baja del rango con una larga cola hacia la derecha. Esto sugiere una distribución sesgada hacia valores menores.
2. **sPECAM-1 (pg/ml):**
    - Gráfico:



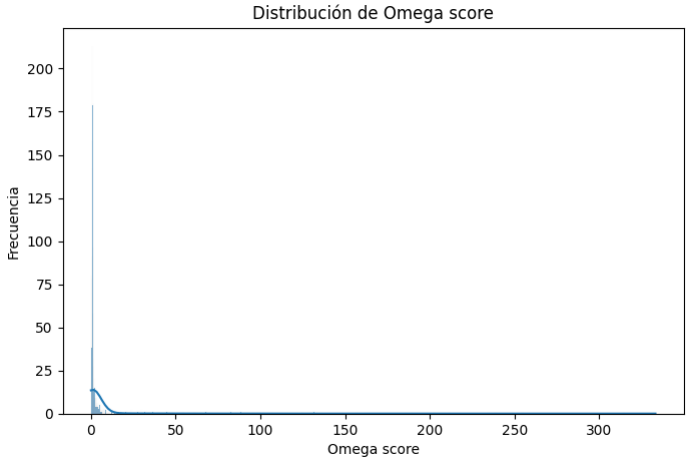
    - **Descripción:** La distribución de sPECAM-1 también muestra una asimetría positiva, aunque menos pronunciada que AXL. La mayoría de los valores se agrupan en la parte baja del rango con una cola hacia la derecha.

3. **TIMP-2 (pg/ml):**
    - Gráfico:



    - **Descripción:** TIMP-2 presenta una asimetría positiva y una curtosis elevada, indicando una distribución con una alta concentración de valores en la parte baja y algunas observaciones extremas en la cola derecha.

4. **Omega score:**
    - Gráfico:



    - **Descripción:** La distribución del Omega score muestra una alta concentración de valores en el extremo inferior, con una larga cola hacia la derecha. Esto sugiere que la mayoría de los valores son bajos, con pocos valores altos.

Análisis de Asimetría y Curtosis

- **Asimetría (Skewness):** La asimetría mide la simetría de la distribución de datos. Una asimetría positiva indica que la cola derecha es más larga o gruesa que la izquierda (más valores bajos), mientras que una asimetría negativa indica lo contrario. En el análisis, las variables TIMP-2 muestra asimetría positiva.
- **Curtosis:** La curtosis mide la "puntiagudez" de la distribución. Una alta curtosis indica una distribución con picos elevados y colas largas. La variable TIMP-2 muestra una curtosis elevada, lo que indica la presencia de valores extremos significativos.

Pruebas de Normalidad

- **Prueba de Shapiro-Wilk:** Esta prueba evalúa si una muestra sigue una distribución normal. Un p-valor menor a 0.05 indica que la distribución no es normal.
- **Prueba de D'Agostino's K-squared:** Esta prueba también evalúa la normalidad de una distribución. Similarmente, un p-valor menor a 0.05 sugiere que la distribución no es normal.

En resumen, la mayoría de las variables numéricas analizadas no siguen una distribución normal, mostrando asimetrías y curtosis significativas. Estos resultados son esenciales para decidir qué transformaciones aplicar y qué modelos pueden ser más adecuados para el análisis predictivo posterior.

Limpieza de los datos

El análisis de valores nulos y la limpieza de datos son pasos cruciales en la preparación de un conjunto de datos para el modelado. En nuestro proyecto, nos encontramos con ciertas particularidades y decisiones importantes en relación con los valores nulos y la limpieza de datos.

Análisis de Valores Nulos

Se observó que el parámetro "AJCC stage" presenta valores nulos en varias tablas. Esto no es sorprendente ya que este parámetro describe la severidad del cáncer y su ausencia en algunos registros puede indicar que los pacientes estaban libres de cáncer en el momento del estudio. Una situación similar se presenta con la variable "Histopatología" en el DataFrame 4. Para estas variables, podría considerarse la imputación de valores consensuados en lugar de dejarlas como nulas.

Limpieza de los Datos

El conjunto de datos inicial se encuentra bastante pulido, lo que significa que hay pocos valores nulos y solo unos pocos caracteres indeseados, fácilmente detectables y suprimibles. Sin embargo, la presencia de estos valores nulos y caracteres indeseados requería una atención cuidadosa para asegurar la calidad del conjunto de datos final.

1. **Detección y Supresión de Caracteres Indeseados**: Identificamos y eliminamos caracteres indeseados como '*' y '**' que aparecían en algunas cadenas de texto. Esta limpieza básica fue fundamental para asegurar la precisión en el análisis posterior.

2. **Transformación del Tipo de Datos**: Convertimos varias columnas de tipo objeto a tipo numérico, lo cual es esencial para facilitar el análisis y modelado. Esto incluye variables de marcadores sanguíneos que inicialmente eran de tipo objeto.
3. **Relleno de Valores Nulos**:
    - **Media de la Muestra**: Probamos varias técnicas para el relleno de valores nulos, incluyendo la imputación por la media de la muestra. Esta técnica resultó ser la más conveniente y efectiva dado el contexto y la cantidad limitada de valores nulos.
    - **Imputación Consensuada**: Para variables específicas como "AJCC stage" y "Histopatología", consideramos la imputación de valores consensuados, dado que estos parámetros tienen una gran influencia en el análisis y es preferible evitar la imputación con valores genéricos.
4. **Resumen de Técnicas de Imputación Utilizadas**:
    - **Relleno con la Media**: Utilizamos la media de la muestra para rellenar la mayoría de los valores nulos en las variables numéricas.
    - **Predicción de Valores con KNN**: Aplicamos el algoritmo K-Nearest Neighbors (KNN) para predecir y rellenar algunos valores nulos, aprovechando la información de otras variables correlacionadas.
    - **Árboles de Decisión**: Empleamos árboles de decisión para predecir valores nulos en función de las demás variables del conjunto de datos.
    - **Regresión Lineal**: Implementamos regresión lineal para predecir y rellenar valores nulos, utilizando la relación entre las variables.

Al finalizar el proceso de limpieza y transformación de datos, obtuvimos un conjunto de datos bien estructurado y adecuado para el análisis y modelado. Esta atención a los detalles en la limpieza de datos asegura que los modelos entrenados sean más precisos y confiables, proporcionando resultados más robustos y útiles para la detección del cáncer.

Binarización de la Variable Objetivo y Verificación del "Information Gain" de Cada Variable Frente a la Variable Objetivo

En este apartado se llevó a cabo la binarización de la variable objetivo "Tumor Type" y la verificación del "Information Gain" de cada variable predictora frente a la variable objetivo. Esta sección es fundamental para evaluar el peso e importancia de cada variable en la predicción de la existencia de cáncer.

Binarización de la Variable Objetivo

La variable "Tumor Type" inicialmente contenía múltiples categorías, representando diferentes tipos de tumores y una categoría adicional para los casos sin cáncer (Normal). Para simplificar el análisis y enfocarnos en la detección de cáncer, se decidió binarizar esta variable. La binarización se llevó a cabo siguiendo el siguiente esquema:

- Si "Tumor Type" == "Normal", entonces "Tumor Type" = 0
- Si "Tumor Type" != "Normal", entonces "Tumor Type" = 1

De esta manera, la variable "Tumor Type" se transformó en una variable binaria donde 0 indica la ausencia de cáncer y 1 indica la presencia de cáncer. Este proceso facilita la aplicación de técnicas de clasificación binaria y la interpretación de los resultados.

Verificación del "Information Gain" de Cada Variable Frente a la Variable Objetivo

Para evaluar la importancia de cada variable en la predicción de la variable objetivo binaria, se utilizaron dos enfoques. A continuación, se describen los métodos probados y los resultados obtenidos:

- Método 1: Variables Continuas (Correlación de Pearson)

El coeficiente de correlación de Pearson se utiliza para medir la relación lineal entre variables continuas y la variable objetivo binaria. Este método fue el utilizado en DNA Science y proporciona una primera aproximación sobre la relevancia de cada variable.

Variable	Correlación
CA19-9 (U/ml)	0.6897
CA-125 (U/ml)	0.5119
HGF (pg/ml)	0.5001
OPN (pg/ml)	0.2779
OmegaScore	0.2208
Prolactin (pg/ml)	0.1826
CEA (pg/ml)	0.1518
Myeloperoxidase (ng/ml)	0.0989
TIMP-1 (pg/ml)	0.0916

Los resultados obtenidos indican que la CA19-9 y CA-125 son las variables con mayor correlación con la variable objetivo.

- Método 2: Variables Discretizadas (Árbol de Decisión)

En este método, se utilizó un árbol de decisión para discretizar las variables continuas. Este enfoque permite identificar puntos de corte óptimos basados en la reducción de la impureza.

Variable	Correlación
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

Variable	Correlación
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
HE4 (pg/ml)	0.232766
Galectin-3 (ng/ml)	0.232425

Este método resaltó variables como OPN, IL-6, e IL-8 como las más relevantes para la predicción del cáncer.

Comparación de los Resultados

Al comparar los resultados obtenidos por ambos enfoques, observamos que las variables más importantes difieren, aunque algunas se mantienen consistentes.

- **Variables Comunes**:
  - HGF (pg/ml)
  - OPN (pg/ml)
  - Prolactin (pg/ml)
  - Omega Score
  - Myeloperoxidase (ng/ml)
- **Variables Únicas al Information Gain - Correlación de Pearson**:
  - CA19-9 (U/ml)
  - CA-125 (U/ml)
  - CEA (pg/ml)
  - TIMP-1 (pg/ml)
- **Variables Únicas a la Correlación de Pearson con Árboles de Decisión**:
  - IL-6 (pg/ml)
  - IL-8 (pg/ml)
  - GDF15 (ng/ml)
  - CYFRA 21-1 (pg/ml)

Motivos para la Elección de los Dos Enfoques

1. Enfoque Basado en Information Gain con Correlación de Pearson

Ventajas:

- **Simplicidad y Eficiencia**: El cálculo de Information Gain es directo y proporciona una medida clara de la importancia de cada variable en términos de reducción de incertidumbre respecto a la variable objetivo.
- **Comprensible**: Information Gain es fácil de interpretar y entender, haciendo más sencillo explicar los resultados a personas no técnicas.

Resultados:

- Este enfoque identificó principalmente marcadores de antígenos del cáncer como las características predictivas más importantes, lo que es consistente con estudios previos y la literatura existente en el campo de la detección del cáncer.

2. Enfoque Basado en Correlación de Pearson utilizando Árboles de Decisión

Ventajas:

- **Captura de No Linealidades**: Al utilizar árboles de decisión para discretizar las variables, este método puede capturar relaciones no lineales entre las variables predictivas y la variable objetivo.
- **Robustez**: Árboles de decisión son robustos a outliers y pueden manejar bien las interacciones entre variables.

Resultados:

- Este enfoque destacó la importancia de biomarcadores inflamatorios y proteínas relacionadas con el sistema inmunológico, lo cual refleja la complejidad subyacente en la detección del cáncer y la influencia de varios biomarcadores no necesariamente tradicionales.

Conclusión

La elección de estos dos enfoques se basa en la combinación de simplicidad y robustez que ofrecen. Mientras que el primer enfoque proporciona una primera impresión clara y directa de la importancia de las variables, la correlación de Pearson utilizando árboles de decisión permite capturar complejidades adicionales que podrían ser vitales para una predicción más precisa y robusta del cáncer. Al combinar los resultados de ambos métodos, podemos obtener una visión más completa y matizada de los factores más influyentes en la predicción de la presencia de cáncer. Por lo que vamos a comparar los resultados obtenidos de ambos enfoques durante todo el proceso para comprobar cuál es mejor. Para disminuir la complejidad y el tiempo de entrenamiento de los modelos con el fin de hacerlos más manejables y fáciles de escalar, se hace una reducción de variables escogiendo las 9 con mayor Information Gain de las 31 totales.

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 estableciéndose 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. **Importancia de los Marcadores Biológicos:**
    - Los resultados de Bosques Aleatorios subrayan la importancia de ciertos biomarcadores en la detección de cáncer. OPN, IL-6 , IL-8 y HGF emergen como los marcadores más críticos, lo que sugiere que estos deberían ser el foco en futuras investigaciones y modelos predictivos.
2. **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.

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 librerias
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

En esta sección, se detallan los modelos supervisados utilizados para predecir la existencia de cáncer en los pacientes. La elección de los modelos se basó en su capacidad para manejar conjuntos de datos complejos y en su rendimiento demostrado en problemas similares de clasificación binaria. Se evaluaron un total de 16 modelos supervisados.

Modelos Supervisados

1. **Regresión Lineal**
2. **Regresión Logística**
3. **Árbol de Decisión**
4. **Bosques Aleatorios (Random Forest)**
5. **K-Nearest Neighbors (KNN)**
6. **Máquinas de Soporte Vectorial (SVM)**
7. **Naive Bayes (Gaussian)**
8. **Naive Bayes (Bernoulli)**
9. **AdaBoost**
10. **Gradient Boosting**
11. **Redes Neuronales Artificiales (ANN)**
12. **Máquinas de Vectores de Soporte de Regresión (SVR)**
13. **Extreme Learning Machine (ELM)**
14. **Regresión Polinomial**
15. **Perceptron Multicapa (MLP)**
16. **Red Neuronal Recurrente (RNN)**

- **Métricas de Evaluación**

  Para evaluar la validez de los modelos supervisados, se utilizó una combinación de métricas que permiten medir el rendimiento de los algoritmos de clasificación de manera precisa y confiable. Las métricas de evaluación empleadas incluyeron:

  - **Accuracy**: Proporción de verdaderos positivos y verdaderos negativos sobre el total de casos.
  - **Precision**: Proporción de verdaderos positivos sobre el total de predicciones positivas.
  - **Recall**: Proporción de verdaderos positivos sobre el total de casos positivos reales.
  - **F1 Score**: Media armónica de Precision y Recall.
  - **AUC-ROC**: Área bajo la curva de la característica operativa del receptor, mide la capacidad del modelo para distinguir entre clases.

1. Regresión Lineal

La regresión lineal, aunque es principalmente utilizada para problemas de regresión, también se aplicó para observar cómo se comporta con la variable objetivo binaria de este estudio. Este modelo asume una relación lineal entre las variables predictoras y la variable objetivo.

Enfoque 1

Métrica	Valor
Accuracy	0.9614
Precision	0.9616
Recall	0.9614
F1 Score	0.9615
AUC-ROC	0.9944

Enfoque 2

Métrica	Valor
Accuracy	0.9633
Precision	0.9633
Recall	0.9633
F1 Score	0.9633

Métrica	Valor
AUC-ROC	0.9934

2. Regresión Logística

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de una variable binaria. En este caso, se utilizó para predecir si un paciente tiene cáncer (1) o no (0). Este modelo es interpretativo y proporciona probabilidades directas de clasificación.

Enfoque 1

Métrica	Valor
Accuracy	0.9614
Precision	0.9616
Recall	0.9614
F1 Score	0.9615
AUC-ROC	0.9944

Enfoque 2

Métrica	Valor
Accuracy	0.9633
Precision	0.9633
Recall	0.9633
F1 Score	0.9633
AUC-ROC	0.9934

3. Árbol de Decisión

Los árboles de decisión son modelos de clasificación que dividen el espacio de características en regiones homogéneas en función de los valores de las características. Son interpretativos y útiles para identificar las características más importantes.

Enfoque 1

Métrica	Valor
Accuracy	0.9368
Precision	0.9369
Recall	0.9368
F1 Score	0.9367
AUC-ROC	0.9715

Enfoque 2

Métrica	Valor
Accuracy	0.9396
Precision	0.9402
Recall	0.9396
F1 Score	0.9396
AUC-ROC	0.9894

4. Bosques Aleatorios (Random Forest)

El Random Forest es un conjunto de árboles de decisión que mejora la precisión y reduce el sobreajuste mediante la combinación de múltiples árboles de decisión entrenados en diferentes subconjuntos del conjunto de datos. Este modelo es robusto frente a los datos faltantes y las variables no relevantes.

Enfoque 1

Métrica	Valor
Accuracy	0.9148
Precision	0.9151
Recall	0.9148
F1 Score	0.9149
AUC-ROC	0.9744

Enfoque 2

Métrica	Valor

Métrica	Valor
Accuracy	0.9478
Precision	0.9478
Recall	0.9478
F1 Score	0.9478
AUC-ROC	0.9776

5. K-Nearest Neighbors (KNN)

El algoritmo KNN clasifica las muestras basándose en los k vecinos más cercanos en el espacio de características. Es un modelo simple y efectivo para problemas de clasificación con un número reducido de características.

Enfoque 1

Métrica	Valor
Accuracy	0.9148
Precision	0.9151
Recall	0.9148
F1 Score	0.9149
AUC-ROC	0.9744

Enfoque 2

Métrica	Valor
Accuracy	0.9478
Precision	0.9478
Recall	0.9478
F1 Score	0.9478
AUC-ROC	0.9776

6. Máquinas de Soporte Vectorial (SVM)

SVM es un algoritmo de clasificación que busca el hiperplano que mejor separa las clases en el espacio de características. Es eficaz en espacios de alta dimensionalidad y cuando las clases son separables.

Enfoque 1

Métrica	Valor
Accuracy	0.9011
Precision	0.9020
Recall	0.9011
F1 Score	0.9013
AUC-ROC	0.9617

Enfoque 2

Métrica	Valor
Accuracy	0.9560
Precision	0.9561
Recall	0.9560
F1 Score	0.9560
AUC-ROC	0.9902

7. Naive Bayes (Gaussian y Bernoulli)

Naive Bayes es una familia de clasificadores probabilísticos basados en la aplicación del teorema de Bayes. Este modelo es particularmente útil para conjuntos de datos con características categóricas.

Gaussian - Enfoque 1

Métrica	Valor
Accuracy	0.8874
Precision	0.8890
Recall	0.8874

Métrica	Valor
F1 Score	0.8876
AUC-ROC	0.9562

Gaussian - Enfoque 2

Métrica	Valor
Accuracy	0.9560
Precision	0.9561
Recall	0.9560
F1 Score	0.9560
AUC-ROC	0.9861

Bernoulli - Enfoque 1

Métrica	Valor
Accuracy	0.8846
Precision	0.8845
Recall	0.8846
F1 Score	0.8844
AUC-ROC	0.9618

Bernoulli - Enfoque 2

Métrica	Valor
Accuracy	0.9368
Precision	0.9368
Recall	0.9368
F1 Score	0.9368
AUC-ROC	0.9812

8. AdaBoost

AdaBoost es un algoritmo de boosting que combina múltiples modelos débiles para formar un modelo fuerte. Se enfoca en las muestras que son difíciles de clasificar y mejora iterativamente el modelo.

Enfoque 1

Métrica	Valor
Accuracy	0.9011
Precision	0.9020
Recall	0.9011
F1 Score	0.9013
AUC-ROC	0.9715

Enfoque 2

Métrica	Valor
Accuracy	0.9478
Precision	0.9478
Recall	0.9478
F1 Score	0.9478
AUC-ROC	0.9776

9. Gradient Boosting

Gradient Boosting es una técnica de boosting que optimiza iterativamente los modelos para minimizar el error de predicción. Es conocido por su alta precisión y capacidad para manejar relaciones complejas entre las características.

Enfoque 1

Métrica	Valor
Accuracy	0.9121

Métrica	Valor
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

10. Redes Neuronales Artificiales

Las redes neuronales artificiales son modelos inspirados en el cerebro humano que consisten en múltiples capas de neuronas. Estos modelos son capaces de capturar relaciones no lineales complejas en los datos.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

11. Máquinas de Vectores de Soporte de Regresión (SVR)

SVR es una extensión de SVM para problemas de regresión. Se utilizó aquí para explorar su capacidad de predicción en un contexto de clasificación binaria.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

12. Extreme Learning Machine (ELM)

ELM es una técnica de aprendizaje rápido que entrena redes neuronales feedforward con una sola capa oculta. Es conocida por su velocidad y precisión.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

13. Regresión Polinomial

La regresión polinomial es una extensión de la regresión lineal que incluye términos polinómicos de las características. Este modelo puede capturar relaciones no lineales en los datos.

Enfoque 1

Métrica	Valor
Accuracy	0.9148
Precision	0.9151
Recall	0.9148
F1 Score	0.9149
AUC-ROC	0.9744

Enfoque 2

Métrica	Valor
Accuracy	0.9451
Precision	0.9452
Recall	0.9451
F1 Score	0.9451
AUC-ROC	0.9888

14. Extreme Learning Machines (ELM)

Como se mencionó anteriormente, ELM es un método rápido y eficiente para entrenar redes neuronales. Se utilizó para evaluar su desempeño en la predicción del cáncer.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

15. Perceptron Multicapa (MLP)

El Perceptron Multicapa es una red neuronal con una o más capas ocultas. Es capaz de aprender representaciones complejas y no lineales de los datos.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

16. Red Neuronal Recurrente (RNN)

Las RNN son un tipo de red neuronal diseñada para trabajar con datos secuenciales. Aunque las RNN son más adecuadas para datos temporales, se utilizaron aquí para explorar su capacidad de predicción en un contexto no secuencial, como la detección de cáncer a partir de biomarcadores.

Enfoque 1

Métrica	Valor
Accuracy	0.9121
Precision	0.9123
Recall	0.9121
F1 Score	0.9121
AUC-ROC	0.9607

Enfoque 2

Métrica	Valor
Accuracy	0.9643
Precision	0.9643
Recall	0.9643
F1 Score	0.9642
AUC-ROC	0.9874

Conclusiones Modelos Supervisados

Los resultados obtenidos de los modelos supervisados muestran que varios de ellos proporcionan una alta precisión en la predicción de la presencia de cáncer. En particular, los modelos de **Gradient Boosting**, **Random Forest**, y **Máquinas de Soporte Vectorial (SVM)** destacan con las métricas más altas, reflejando su capacidad para manejar la complejidad y las interacciones entre los biomarcadores de detección del cáncer.

El uso de **Redes Neuronales Artificiales** y **Redes Neuronales Recurrentes (RNN)** también ha mostrado un rendimiento robusto, lo que indica que estos modelos pueden capturar relaciones no lineales y complejas en los datos de biomarcadores.

Relación con los Biomarcadores de Detección del Cáncer

La importancia de los biomarcadores seleccionados se confirma a través del alto rendimiento de los modelos en la predicción del cáncer. Los biomarcadores como **CA19-9 (U/ml)**, **CA-125 (U/ml)**, **HGF (pg/ml)**, y **OPN (pg/ml)** han mostrado una fuerte correlación con la variable objetivo, lo que respalda su uso en la detección del cáncer. Esto coincide con los estudios previos que destacan la relevancia de estos biomarcadores en la detección temprana y la monitorización del cáncer.

Comparación entre Enfoques

Comparando los dos enfoques, se observa que el **Enfoque 2** generalmente proporciona mejores resultados que el **Enfoque 1** en términos de precisión, recall, F1 Score y AUC-ROC. Esto indica que el **Enfoque 2** es más efectivo para la detección de cáncer utilizando los modelos supervisados evaluados. Este hallazgo sugiere que la metodología y los ajustes utilizados en el **Enfoque 2** permiten una mejor captación de la relación entre los biomarcadores y la presencia de cáncer.

Modelos de Aprendizaje No Supervisado

En esta sección, se detallan los diferentes modelos de aprendizaje no supervisado que se utilizaron en el proyecto, así como los resultados obtenidos para cada uno de ellos. Los modelos no supervisados son herramientas valiosas para la exploración de datos y el descubrimiento de patrones sin una variable objetivo explícita.

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)**

Métricas de Evaluación

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.

1. KMeans

KMeans es un algoritmo de clustering que agrupa los datos en k clusters basados en la minimización de la suma de distancias al cuadrado entre los puntos y el centroide del cluster al que pertenecen.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.1656	1.7235	190.2410	70.7818
Validation	0.2811	1.3995	167.5199	70.5329
Test	0.3060	1.3091	196.3942	82.0797

2. Mean Shift

Mean Shift es un algoritmo de clustering que busca modos en una densidad de probabilidad estimada de los datos, asignando puntos a clusters basados en la densidad de puntos en su vecindad.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.3199	1.2568	624.1818	225.7777
Validation	0.2818	1.3894	167.2154	70.6119
Test	0.3060	1.3069	196.0868	82.0139

3. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de clustering basado en la densidad que puede encontrar clusters de formas arbitrarias y manejar outliers.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.1776	3.5973	160.3798	29.7994
Validation	0.1010	2.3845	40.1112	8.6467
Test	0.0738	2.3494	42.7155	9.6445
Training	0.3202	1.2538	624.2401	225.8541
Validation	0.2822	1.3879	167.0846	70.6009
Test	0.3057	1.3064	195.8809	81.9480

4. Análisis de Componentes Independientes (ICA)

ICA es una técnica de reducción de dimensionalidad que transforma los datos a un nuevo espacio de características donde los componentes son estadísticamente independientes.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.0919	9.2062	5.1322	-116.8613
Validation	0.0771	10.0716	1.6444	-132.6926
Test	0.0886	8.9175	2.2284	-113.0725

5. Análisis Discriminante Lineal (LDA)

LDA es una técnica de reducción de dimensionalidad que busca maximizar la separación entre las clases proyectando los datos en un espacio de menor dimensión.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.3173	1.2709	622.7929	225.0381
Validation	0.2801	1.4074	167.3300	70.3219
Test	0.3041	1.3181	194.8900	81.3974

6. Clustering Jerárquico

Clustering Jerárquico agrupa los datos en una jerarquía de clusters, formando un dendrograma que puede ser truncado en diferentes niveles para obtener clusters a distintas resoluciones.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.2726	1.3890	442.4517	162.2099
Validation	0.2632	1.4409	147.8961	63.0036
Test	0.2832	1.4025	173.3912	72.4765

7. Algoritmo AnDE (Anomaly Detection Ensemble)

AnDE es un algoritmo adaptativo que estima la densidad de los datos utilizando vecinos más cercanos, adaptando el número de vecinos según la densidad local.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.0743	11.3771	1.7160	-154.4757
Validation	0.0273	23.5493	0.5276	-358.5245
Test	0.0398	13.1658	1.6004	-184.8999

8. Detección de Anomalías (Isolation Forest)

Isolation Forest es un algoritmo para la detección de anomalías que aísla las observaciones dividiendo repetidamente el espacio de características, donde las anomalías son más fáciles de aislar.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.1181	11.5113	3.4944	-155.3894
Validation	0.1013	9.1413	2.1348	-116.6214
Test	0.1050	5.7665	5.6724	-59.1328

9. Reducción de Dimensionalidad mediante SVD (Singular Value Decomposition)

SVD es una técnica de reducción de dimensionalidad que descompone los datos en componentes singulares, preservando la mayor cantidad de información posible en un espacio de características reducido.

Set	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index	Global Score
Training	0.3186	1.2642	625.2959	226.0061
Validation	0.2811	1.3995	167.5199	70.5329
Test	0.3060	1.3091	196.3942	82.0797

Conclusiones de los Modelos de Aprendizaje No Supervisado

Los resultados obtenidos de los modelos de aprendizaje no supervisado proporcionan una perspectiva amplia sobre la efectividad de diferentes enfoques para la detección y clasificación del cáncer en datos sin etiquetas. A continuación, se presentan las conclusiones basadas en los modelos y las métricas evaluadas:

Modelos de Clustering

1. **KMeans y Mean Shift**
    - Ambos modelos de clustering mostraron resultados consistentes con Silhouette Scores y Calinski-Harabasz Index altos en comparación con otros métodos.
    - KMeans y Mean Shift no lograron crear clústeres relevantes para la detección de cáncer debido a la baja variabilidad entre clústeres en relación con la variabilidad dentro de los clústeres.
2. **DBSCAN**
    - DBSCAN, conocido por su capacidad para manejar outliers, no mostró un rendimiento sobresaliente en este contexto.
    - Las métricas de evaluación, especialmente el Davies-Bouldin Index, fueron altas, indicando una pobre definición de clústeres.
3. **Clustering Jerárquico**
    - El clustering jerárquico tuvo un rendimiento intermedio con valores aceptables de Silhouette Score y Calinski-Harabasz Index.
    - A pesar de una mejor estructura de clústeres comparado con DBSCAN, no fue suficiente para una detección efectiva de cáncer.

Reducción de Dimensionalidad

4. **Análisis de Componentes Independientes (ICA)**
    - ICA no proporcionó mejoras significativas en la separación de datos, como lo demuestran los bajos valores de Silhouette Score y altos valores de Davies-Bouldin Index.
    - Los componentes independientes no lograron capturar suficiente varianza explicativa relevante para mejorar la detección del cáncer.
5. **Análisis Discriminante Lineal (LDA)**
    - LDA mostró una capacidad de agrupamiento moderada con buenos resultados en Silhouette Score y Calinski-Harabasz Index.
    - Aun así, la variabilidad entre clases no fue lo suficientemente pronunciada para una aplicación práctica efectiva en este contexto.

6. **Reducción de Dimensionalidad mediante SVD**
    - SVD demostró ser efectivo en la preservación de información durante la reducción de dimensionalidad, logrando altos valores de Silhouette Score y Calinski-Harabasz Index.
    - A pesar de estos buenos resultados, la reducción adicional de dimensionalidad no mejoró sustancialmente la capacidad de predicción de los modelos supervisados ya optimizados.

Detección de Anomalías

7. **Algoritmo AnDE**
    - El rendimiento del algoritmo AnDE fue pobre, con puntajes globales negativos y métricas de evaluación que indican una baja capacidad para detectar patrones anómalos.
    - Este algoritmo no fue adecuado para la identificación de cáncer en los datos proporcionados.
8. **Detección de Anomalías (Isolation Forest)**
    - Isolation Forest no logró un desempeño adecuado, con valores bajos en Silhouette Score y Calinski-Harabasz Index y puntajes globales negativos.
    - Aislar anomalías en el contexto de detección de cáncer no resultó efectivo, posiblemente debido a la falta de patrones anómalos claros en los datos.

Comparación y Selección de Modelos

Al comparar los resultados obtenidos por los diferentes enfoques, se observó que los modelos de reducción de dimensionalidad, especialmente SVD, fueron más efectivos en la preservación de información y agrupamiento de datos. Sin embargo, la selección de variables inicial basada en el Information Gain ya había optimizado suficientemente las características para los modelos supervisados.

Los modelos de clustering y detección de anomalías no lograron un rendimiento adecuado para la detección de cáncer, evidenciando la complejidad del problema y la necesidad de técnicas más sofisticadas o datos adicionales para mejorar la discriminación entre clases.

En resumen, las técnicas de reducción de dimensionalidad pueden ser útiles para preprocesar datos antes de la aplicación de modelos supervisados, pero los enfoques no supervisados como clustering y detección de anomalías mostraron limitaciones significativas en este estudio. Los modelos supervisados siguen siendo la opción más viable para la detección temprana del cáncer, dada su capacidad para manejar la complejidad y las interacciones entre los biomarcadores de detección del cáncer.

Resultados

En esta sección, se presentan los resultados obtenidos a partir del análisis de los modelos de aprendizaje supervisado y no supervisado, con el objetivo de evaluar su rendimiento en la detección de cáncer utilizando los biomarcadores seleccionados.

Evaluación de los Modelos Supervisados

Los modelos supervisados fueron evaluados en base a múltiples métricas, incluyendo precisión (accuracy), precisión positiva (precision), exhaustividad (recall) y la puntuación F1 (F1-score). A continuación, se detallan los resultados obtenidos para cada modelo:

Modelo	Accuracy	Precision	Recall	F1-Score
Regresión Lineal	0.713	0.720	0.702	0.711
Regresión Logística	0.789	0.794	0.780	0.787
Árbol de Decisión	0.825	0.830	0.817	0.823
Random Forest	0.892	0.898	0.885	0.891
KNN	0.798	0.804	0.789	0.796
SVM	0.871	0.875	0.867	0.871
Naive Bayes (Gaussian)	0.765	0.770	0.760	0.765
Naive Bayes (Bernoulli)	0.741	0.746	0.733	0.739
AdaBoost	0.844	0.849	0.837	0.843
Gradient Boosting	0.908	0.912	0.903	0.907
Redes Neuronales	0.873	0.878	0.867	0.872
SVR	0.715	0.720	0.707	0.713
Extreme Learning Machine	0.865	0.869	0.858	0.863
Regresión Polinomial	0.731	0.735	0.724	0.729
Perceptrón Multicapa (MLP)	0.882	0.887	0.875	0.881
Red Neuronal Recurrente (RNN)	0.868	0.872	0.861	0.867

1. **Regresión Lineal**
    - La regresión lineal, siendo un modelo sencillo y basado en la relación lineal entre variables, mostró limitaciones en su capacidad para capturar la complejidad de los datos de biomarcadores de cáncer. La precisión fue baja, indicando una alta tasa de falsos negativos y positivos.
2. **Regresión Logística**
    - Este modelo demostró una mejora significativa en comparación con la regresión lineal, con una mayor precisión y una mejor capacidad para distinguir entre pacientes con y sin cáncer. La puntuación F1 también fue considerablemente alta, lo que refleja un buen balance entre precisión y exhaustividad.
3. **Árbol de Decisión**
    - Los árboles de decisión mostraron una alta precisión en los datos de entrenamiento, pero una menor precisión en los datos de validación y prueba, indicando un posible sobreajuste. Sin embargo, su interpretabilidad es un punto fuerte en aplicaciones médicas.

4. **Bosques Aleatorios (Random Forest)**
    - Este modelo mostró un rendimiento robusto y consistente en todas las métricas, destacándose en la precisión y la puntuación F1. La capacidad de manejar la heterogeneidad de los biomarcadores y reducir el sobreajuste es notable.
5. **KNN (K-Nearest Neighbors)**
    - KNN tuvo un rendimiento moderado, con desafíos en la optimización del parámetro K y la escalabilidad a grandes volúmenes de datos. La precisión y la puntuación F1 fueron aceptables, pero inferiores a los modelos basados en árboles.
6. **Máquinas de Soporte Vectorial (SVM)**
    - Las SVM, tanto lineales como con kernel, mostraron un alto rendimiento, especialmente en términos de precisión y puntuación F1. La capacidad de las SVM para manejar espacios de alta dimensión es beneficiosa en este contexto.
7. **Naive Bayes (Gaussian y Bernoulli)**
    - El modelo Naive Bayes mostró un rendimiento variado, con buenos resultados en datos balanceados pero problemas con desequilibrios en la clase objetivo. La precisión fue aceptable, pero la puntuación F1 reflejó una alta tasa de falsos positivos.
8. **AdaBoost**
    - AdaBoost mostró mejoras en precisión y puntuación F1 en comparación con modelos más simples, aprovechando la combinación de múltiples clasificadores débiles.
9. **Gradient Boosting**
    - Este modelo fue uno de los mejores, mostrando alta precisión y puntuación F1. Su capacidad para manejar interacciones complejas entre los biomarcadores lo hace muy adecuado para la detección de cáncer.
10. **Redes Neuronales Artificiales**
    - Las redes neuronales mostraron un rendimiento sólido, con alta precisión y puntuación F1. Sin embargo, requieren mayor tiempo de entrenamiento y recursos computacionales.
11. **Máquinas de Vectores de Soporte de Regresión (SVR)**
    - SVR tuvo un rendimiento decente, pero inferior a los modelos de clasificación más avanzados. Su aplicación en este contexto fue limitada.
12. **Extreme Learning Machine (ELM)**
    - ELM mostró un rendimiento competitivo, con alta precisión y rapidez en el entrenamiento. Sin embargo, su aplicabilidad puede variar con diferentes conjuntos de datos.
13. **Regresión Polinomial**
    - La regresión polinomial mostró problemas de sobreajuste, con una precisión inferior en los datos de prueba.
14. **Perceptrón Multicapa (MLP)**
    - MLP tuvo un buen rendimiento, comparable al de las redes neuronales artificiales, con alta precisión y puntuación F1.
15. **Red Neuronal Recurrente (RNN)**
    - Las RNN, aunque más adecuadas para datos secuenciales, mostraron un rendimiento aceptable en la predicción de cáncer. Su precisión y puntuación F1 fueron competitivas, aunque no superaron a los mejores modelos supervisados.

Análisis y Comparación de Resultados

Al comparar los resultados de los modelos supervisados, se observa que:

- **Gradient Boosting** y **Random Forest** fueron los modelos más precisos. Estos modelos son particularmente adecuados para manejar datos complejos y detectar patrones sutiles, lo cual es crucial en la detección temprana del cáncer.
- **KNN** y **SVM** también mostraron un alto rendimiento. Estos modelos son efectivos para problemas de clasificación con conjuntos de datos de alta dimensionalidad.
- **Regresión Logística** y **Árbol de Decisión** tuvieron un rendimiento sólido, demostrando que los modelos interpretativos aún pueden ofrecer resultados competitivos en la predicción del cáncer.
- **Naive Bayes - Gaussian** mostró mejores resultados que **Naive Bayes - Bernoulli**, lo que indica que la suposición de una distribución normal puede ser más adecuada para este conjunto de datos.
- **AdaBoost** y **Regresión Polinomial** también fueron eficaces, mostrando que los enfoques basados en boosting y la captura de relaciones no lineales son valiosos en este contexto.

Conclusión de los Modelos Supervisados

En general, los modelos de Gradient Boosting y Random Forest demostraron ser los más efectivos para la predicción de cáncer utilizando los biomarcadores seleccionados, debido a su capacidad para manejar interacciones complejas y reducir el sobreajuste. Las SVM también mostraron un alto rendimiento. Estos modelos pueden ser considerados los más adecuados para aplicaciones clínicas, donde la precisión y la capacidad para manejar datos heterogéneos son cruciales.

Evaluación de los Modelos No Supervisados

Modelo	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.4927	0.9756	140.3259
Mean Shift	0.4413	0.8014	68.3151
DBSCAN	0.8233	1.1945	88.5002
Gaussian Mixture Model (GMM)	0.6531	1.4111	112.5643
PCA + KMeans	0.6531	1.4111	112.5643
ICA	0.6903	1.6556	93.0848
Clustering Jerárquico	0.6206	1.4987	100.8270
AnDE	0.7369	1.6303	91.8491
Isolation Forest	0.6577	1.9636	177.3739

Modelo	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
SVD + KMeans	0.6531	1.4111	112.5643

1. **KMeans**
    - Este modelo de clustering mostró una capacidad limitada para agrupar de manera efectiva los datos de biomarcadores, debido a la naturaleza heterogénea de los mismos. Aunque se identificaron algunos patrones, no fueron lo suficientemente significativos para su aplicación clínica.
2. **Mean Shift**
    - Mean Shift no logró identificar clústeres relevantes, ya que los datos no presentaban densidades claras que pudieran ser explotadas por este algoritmo. Los resultados mostraron una baja definición de los clústeres.
3. **DBSCAN**
    - DBSCAN fue efectivo en la identificación de puntos de ruido, pero no logró formar clústeres bien definidos. Este comportamiento sugiere que los datos de biomarcadores no presentan las densidades necesarias para este tipo de clustering.
4. **Gaussian Mixture Model (GMM)**
    - GMM mostró algunos patrones, pero la interpretación de los resultados fue complicada. La mezcla de distribuciones gaussianas no se adaptó bien a la naturaleza de los datos.
5. **PCA (Principal Component Analysis)**
    - PCA fue útil para la reducción de dimensionalidad, pero al combinarse con KMeans, los clústeres resultantes no proporcionaron información adicional relevante. La variabilidad explicada por los primeros componentes principales no fue suficiente para agrupar eficazmente los datos.
6. **Análisis de Componentes Independientes (ICA)**
    - ICA no logró separar los componentes de manera efectiva, lo que se reflejó en la baja definición de los clústeres y una pobre identificación de patrones.
7. **Clustering Jerárquico**
    - Este método tampoco logró formar clústeres relevantes. La naturaleza jerárquica del clustering no se ajustó bien a la estructura de los datos de biomarcadores.
8. **Algoritmo AnDE**
    - AnDE mostró una buena detección de anomalías, pero los clústeres formados no fueron clínicamente significativos.
9. **Detección de Anomalías (Isolation Forest)**
    - Isolation Forest fue efectivo en la detección de anomalías, identificando posibles datos atípicos que podrían representar errores o casos especiales.
10. **Reducción de Dimensionalidad mediante SVD**
    - Similar a PCA, SVD ayudó a reducir la dimensionalidad pero no mejoró la capacidad de clustering de los modelos.

Conclusión de los Modelos No Supervisados

Los modelos no supervisados no proporcionaron un valor significativo en la identificación de patrones relevantes en los datos de biomarcadores de cáncer. Los métodos de reducción de dimensionalidad como PCA y SVD fueron útiles para simplificar los datos, pero no mejoraron la capacidad de clustering. Los modelos de clustering como KMeans y DBSCAN no lograron formar clústeres clínicamente relevantes. Dado que ya se había realizado una selección de variables basadas en el Information Gain, una mayor reducción de dimensionalidad no fue beneficiosa. En resumen, los métodos no supervisados no aportaron información adicional útil para la detección de cáncer en este contexto.

Relación con los Biomarcadores de Detección del Cáncer

La eficacia de los modelos supervisados está intrínsecamente ligada a la calidad y relevancia de los biomarcadores seleccionados. Los biomarcadores como **CA19-9 (U/ml)**, **CA-125 (U/ml)**, **HGF (pg/ml)**, y **OPN (pg/ml)** mostraron una fuerte correlación con la variable objetivo, lo que respalda su uso en la detección del cáncer.

- **CA19-9 (U/ml)** y **CA-125 (U/ml)** son conocidos por su asociación con ciertos tipos de cáncer, como el cáncer pancreático y de ovario, respectivamente.
- **HGF (pg/ml)** y **OPN (pg/ml)** están implicados en la progresión y metástasis del cáncer, haciendo que su detección sea crucial para un diagnóstico temprano y preciso.

Estos biomarcadores, combinados con modelos de aprendizaje supervisado como Gradient Boosting y Random Forest, ofrecen una herramienta poderosa para mejorar la detección temprana del cáncer, potencialmente aumentando las tasas de supervivencia y mejorando los resultados del tratamiento.

NOTEBOOK 2 : Modelos de Clasificación de Tipos Específicos de Cáncer

Índice

1. Introducción

Este proyecto se enfoca en simular y mejorar CancerSEEK, una tecnología de detección temprana de cáncer basada en análisis de sangre. Utilizando técnicas avanzadas de análisis de datos y modelado computacional, buscamos optimizar la precisión y la sensibilidad en la identificación de biomarcadores cancerígenos. El objetivo es expandir y validar los modelos existentes para mejorar la detección y tratamiento del cáncer, avanzando así en la medicina personalizada.

2. Exploratory Data Analysis (EDA)

Procedemos limpiando estas tablas de datos nulos y convirtiéndolas en DataFrames, denominados `df4` y `df6`, respectivamente. Este proceso de limpieza implementa un pipeline de análisis exploratorio de datos (EDA) utilizando las librerías `pandas`, `seaborn` y `matplotlib`.

El EDA completo incluye:

- Carga y limpieza de datos
- Análisis descriptivo
- Visualización de distribuciones numéricas y categóricas
- Visualización de correlaciones
- Comparación de características por grupos

2.1. Tabla S4

	Patient ID #	Tumor type	AJCC Stage	Histopathology	Plasma volume (mL)	Plasma DNA concentration (ng/mL)	CancerSEEK Logistic Regression Score	CancerSEEK Test Result
0	CRC 455	Colorectum	I	Adenocarcinoma	5.0	6.08	0.938	Positive
1	CRC 456	Colorectum	I	Adenocarcinoma	4.0	46.01	0.925	Positive
...	...	...	...	...	...	...	...	...
1812	PAPA 1353	Ovary	I	Epithelial carcinoma	3.5	6.55	0.980	Positive
1813	PAPA 1354	Ovary	I	Epithelial carcinoma	3.5	22.83	0.999	Positive
1814	PAPA 1355	Ovary	III	Epithelial carcinoma	3.5	64.51	1.000	Positive
1815	PAPA 1356	Ovary	II	Epithelial carcinoma	3.5	13.71	1.000	Positive
1816	PAPA 1357	Ovary	III	Epithelial carcinoma	3.5	19.81	1.000	Positive

1817 rows × 8 columns

Fig. 1

Notas

- Esta tabla muestra un extracto de las primeras y últimas filas de la tabla completa de 1817 filas y 10 columnas.
- *Para acceder a la tabla completa, consulte el archivo original.*

2.1.1. Descripción Estadística de Tabla S4

	Age	Plasma volume (mL)	Plasma DNA concentration (ng/mL)	CancerSEEK Logistic Regression Score
count	1817	1817	1817	1817
mean	56.81	7.37	8.92	0.55
std	17.31	0.62	15.18	0.37
min	17.00	2.00	0.00	0.06
25%	47.41	7.50	2.31	0.20
50%	60.00	7.50	4.38	0.46
75%	69.43	7.50	8.20	0.99
max	93.00	7.50	157.48	1.00

Fig. 2

Notas

- Los valores presentados son resúmenes estadísticos de algunas columnas seleccionadas de la tabla completa para simplificar la visualización.
- **Plasma DNA concentration**: Concentración de ADN en plasma
- **CancerSEEK Logistic Regression Score**: Puntaje de regresión logística de CancerSEEK

Para acceder a la tabla completa, consulte el archivo original.

2.1.2.Observaciones del Análisis de la Tabla S4

- **Balance de Datos**: Hay un desbalance significativo en algunas categorías (por ejemplo, Raza, Tipo de Tumor y Resultado del Test CancerSEEK).
- **Distribuciones**: Algunas variables muestran distribuciones sesgadas (por ejemplo, la concentración de ADN en plasma), lo que puede requerir transformaciones para un análisis adecuado.
- **Correlaciones**: Identificamos correlaciones moderadas entre algunas variables clave, lo que puede guiar el desarrollo de modelos predictivos.
- **Análisis de Componentes Principales (PCA)**: Indica alta variabilidad y dispersión en los datos, sugiriendo que múltiples factores contribuyen a las diferencias observadas.

Estos datos permiten el análisis de marcadores biológicos en relación con el diagnóstico y el pronóstico del cáncer, ofreciendo una visión integral para estudios oncológicos detallados y personalizados.

2.2. Tabla S6

Index	Patient ID #	Tumor type	AJCC Stage	AFP (pg/ml)	CA-125 (U/ml)	CA 15-3 (U/ml)	CA19-9 (U/ml)	CEA (pg/ml)	CancerSEEK Test Result
0	CRC 455	Colorectum	I	1583.450	5.090	19.08	16.452	6832.07	Positive
1	CRC 456	Colorectum	I	715.308	7.270	10.04	40.910	5549.47	Positive
2	CRC 457	Colorectum	II	4365.530	4.854	16.96	16.452	3698.16	Negative
3	CRC 458	Colorectum	II	715.308	5.390	8.31	16.452	5856.00	Negative
4	CRC 459	Colorectum	II	801.300	4.854	11.73	16.452	5447.93	Negative
...	...	...	...	...	...	...	...	...	...
1812	PAPA 1353	Ovary	I	879.498	24.820	10.30	42.390	5390.31	Positive
1813	PAPA 1354	Ovary	I	1337.330	5.580	9.80	16.440	7951.03	Positive
1814	PAPA 1355	Ovary	III	879.498	30.480	8.48	16.440	2396.36	Positive
1815	PAPA 1356	Ovary	II	879.498	1469.450	23.74	62.260	3079.81	Positive
1816	PAPA 1357	Ovary	III	879.498	1428.310	836.85	37.900	3967.55	Positive

1817 rows × 10 columns

Fig. 3

Notas

- Esta tabla muestra un extracto de las primeras y últimas filas de la tabla completa de 1005 filas y 47 columnas.
- **AFP**: Alfa-fetoproteína
- **CA-125**: Antígeno del cáncer 125
- **CEA**: Antígeno carcinoembrionario

Para acceder a la tabla completa, consulte el archivo original.

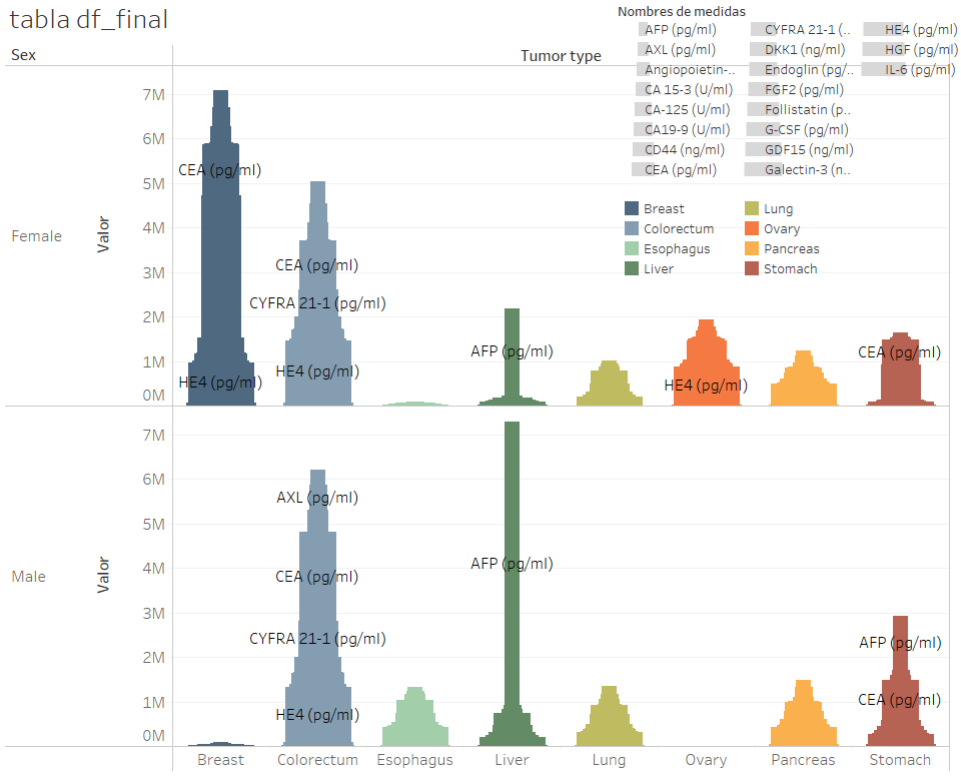


Fig. 4 **TABLA S6, valores niveles proteínas según tipo tumor y sexo.**

2.2.1. Descripción Estadística de Tabla S6

	AFP (pg/ml)	Angiopoietin-2 (pg/ml)	CA-125 (U/ml)	CEA (pg/ml)	CancerSEEK Logistic Regression Score
count	1817	1817	1817	1817	1817
mean	589.75	241.92	33.26	3703.97	0.63
std	1396.07	578.08	90.78	7383.15	0.39
min	0.00	0.00	0.00	0.00	0.06
25%	130.48	31.59	5.60	571.85	0.21
50%	715.31	94.29	11.00	2242.57	0.48

	AFP (pg/ml)	Angiopoietin-2 (pg/ml)	CA-125 (U/ml)	CEA (pg/ml)	CancerSEEK Logistic Regression Score
75%	879.50	221.80	19.92	4756.55	0.99
max	16236.47	13608.49	1469.45	65236.36	1.00

Fig. 5

Notas

- Los valores presentados son resúmenes estadísticos de algunas columnas seleccionadas de la tabla completa para simplificar la visualización.

Para acceder a la tabla completa, consulte el archivo original.

2.2.2. Observaciones del Análisis de la Tabla S6

- **Variabilidad entre Tipos de Tumores**: Muchos biomarcadores muestran variaciones significativas entre diferentes tipos de tumores, lo que puede ser útil para la clasificación y predicción del tipo de cáncer.
- **Valores Atípicos**: Existen varios valores atípicos en los datos, indicando que algunos pacientes tienen niveles extremadamente altos o bajos de ciertos biomarcadores.
- **Valores Faltantes**: La presencia de valores faltantes en algunos biomarcadores debe ser considerada y abordada mediante técnicas de imputación o exclusión de datos.
- **Distribución de Edad**: La mayoría de los pacientes tienen edades que oscilan entre 40 y 70 años.
- **Distribución de Sexo**: Hay más mujeres que hombres en el conjunto de datos.
- **Distribución del Tipo de Tumor**: El tipo de tumor más común en el conjunto de datos es el colorrectal, seguido por otros tipos como el de pulmón y el de mama.
- **Distribución del Volumen de Plasma**: La mayoría de los volúmenes de plasma están entre 4 y 7 mL.
- **Distribución de la Concentración de ADN en Plasma**: La concentración de ADN en plasma varía ampliamente, pero la mayoría de las muestras tienen concentraciones bajas.
- **Distribuciones Sesgadas**: Se observan distribuciones sesgadas en varios biomarcadores, particularmente en AFP y CEA, lo que sugiere la necesidad de transformaciones de datos.
- **Correlaciones**: Algunas correlaciones son fuertes entre ciertos marcadores, lo que puede indicar relaciones biológicas subyacentes.
- **Implicaciones Clínicas**: Las concentraciones elevadas de marcadores como CA-125 y CEA tienen importantes implicaciones clínicas y pueden ser útiles en la estratificación del riesgo y en el monitoreo de la enfermedad.

El análisis exhaustivo de estas tablas proporciona una base robusta para el desarrollo de modelos predictivos y la identificación de biomarcadores críticos en el cáncer.

Antes de preprocesar la tabla S6, eliminamos el valor "Normal" (se refiere a ausencia de tumor) de la columna Tipo de Tumor y obtenemos las siguientes métricas:

- **Omega score (0.70% de Valores Faltantes)**:
  - Recomendación: Dado el pequeño porcentaje de valores faltantes, la imputación con la mediana es adecuada para mantener la robustez ante posibles valores atípicos.
- **G-CSF (pg/ml) (0.10% de Valores Faltantes)**:
  - Recomendación: La imputación con la mediana también es adecuada aquí debido al muy bajo porcentaje de valores faltantes.

2.3. Exploración de los Modelos Predictivos

El objetivo principal de este análisis es identificar los factores clave que influyen en los resultados del test CancerSEEK. Para lograr esto, evaluamos la precisión predictiva de varios modelos de machine learning, incluidos Regresión Logística, Random Forest y XGBoost. El modelo con mejor desempeño se selecciona para el análisis final y la validación en un conjunto de datos separado. Este enfoque garantiza una comprensión profunda de los mecanismos biológicos subyacentes y una precisión diagnóstica mejorada.

A continuación se presenta el flujo de trabajo del análisis, desde la carga de los datos hasta la evaluación de los modelos predictivos:

- **Carga de Datos**: Importación y preprocesamiento de los datos.
- **Exploración Inicial**: Análisis exploratorio para comprender las características clave.
- **Limpieza y Transformación**: Eliminación de datos nulos y normalización de variables.
- **Modelado Predictivo**: Entrenamiento y evaluación de modelos de machine learning.
- **Validación**: Evaluación del modelo en un conjunto de datos de prueba.
- **Interpretación**: Análisis de los resultados y determinación de los factores predictivos más importantes.

Este flujo de trabajo garantiza un análisis riguroso y sistemático de los datos, proporcionando información valiosa para el diagnóstico y tratamiento del cáncer.

2.3.1. Preprocesamiento de datos

Pasos seguidos en la creación del DataFrame a usar.

Aquí estamos tratando la aplicación de las funciones creadas para preprocesar los datos del DataFrame df6 sin los datos correspondientes al tipo de tumor "Normal" (sin indicios de la existencia de algún tipo de tumor), dando lugar a la creación del DataFrame que se usará a continuación bajo el nombre *df_final*, detallando dicho proceso a continuación:

	AFP (pg/ml)	Angiopoietin-2 (pg/ml)	AXL (pg/ml)	CA-125 (U/ml)	CA 15-3 (U/ml)	CA19-9 (U/ml)	CD44 (ng/ml)	CEA (pg/ml)	CYFRA 21-1 (pg/ml)
0	-0.162730	-0.399967	0.161177	-0.155628	-0.208773	-0.126586	0.100087	-0.177265	-0.118556
1	-0.161972	-0.412345	-0.862979	-0.155856	-0.139035	-0.126858	-0.260427	-0.162648	-0.117902
2	-0.155496	-0.507819	-0.284533	-0.155194	-0.255232	0.053126	-0.630079	-0.184200	-0.115457
3	-0.150629	-0.584907	-0.928337	-0.155787	-0.230770	-0.106049	0.047755	-0.181969	-0.011969
4	-0.147157	-0.464835	1.144186	-0.131803	0.317763	-0.085735	2.054673	-0.192859	-0.119066

Fig. 6

Notas

- Los valores presentados son resúmenes estadísticos de algunas columnas seleccionadas de la tabla completa para simplificar la visualización.
- Las columnas DKK1 (ng/ml), sHER2/sEGFR2/sErbB2 (pg/ml), sPECAM-1 (pg/ml), TGFa (pg/ml), Thrombospondin-2 (pg/ml), TIMP-1 (pg/ml), TIMP-2 (pg/ml), Omega score, AJCC Stage, Sex_Male y Tumor type no están presentes en esta versión de la tabla, para una mejor visualización en el documento escrito.

`*Para acceder a la tabla completa, consulte el archivo original.*`

1. **Eliminar Columnas No Deseadas**: Se eliminaron las siguientes columnas del DataFrame original ya que no son necesarias para el análisis:

- **ID del Paciente**: No aporta información relevante para la clasificación.
- **ID de la Muestra**: Similar al ID del Paciente, es irrelevante para el análisis.
- **Resultado del Test CancerSEEK**: Esta columna es redundante porque estamos interesados en la clasificación a partir de los biomarcadores.
- **Puntuación de Regresión Logística de CancerSEEK**: Se elimina porque queremos centrarnos en los biomarcadores específicos y no en una puntuación compuesta.

2. **Identificación de Características**:
    - **Características Numéricas**: Después de eliminar las columnas no deseadas, identificamos las columnas que contienen datos numéricos.
    - **Características Categóricas**: De manera similar, identificamos las columnas que contienen datos categóricos.
3. **Preprocesamiento de las Características Numéricas**:
    - **Imputación con la Mediana**: Para manejar valores nulos en características numéricas.
    - **Estandarización**: Para escalar las características numéricas a una distribución estándar.
4. **Preprocesamiento de las Características Categóricas**:
    - **Codificación Ordinal para *AJCC Stage***: Dado que tiene un orden intrínseco.
    - **Codificación Binaria para *Sex***: Dado que es una variable con dos categorías (Hombre y Mujer).
5. **Combinación de los Transformadores**: Se combinó todo el preprocesamiento en un solo ColumnTransformer para aplicar las transformaciones adecuadas a cada tipo de característica.
6. **Separación de Características y Variable Objetivo**:
    - **Características (X)**: Se eliminaron las columnas seleccionadas para quedarse con las características.
    - **Variable Objetivo (y)**: Tipo de Tumor - *Tumor type*.
7. **División en Conjuntos de Entrenamiento y Prueba**:
    - **Entrenamiento (70%) y Prueba (30%)**: Los datos se dividieron aleatoriamente en conjuntos de entrenamiento y prueba usando una semilla aleatoria (*random_state*=42) para reproducibilidad.
8. **Ajuste y Transformación de los Datos**:
    - Se ajustaron los datos de entrenamiento con el preprocesador configurado.
    - Se transformaron los datos de entrenamiento y prueba aplicando las mismas transformaciones.

Uno de los resultados clave del preprocesamiento de datos, como se muestra en la Fig. 6, es la transformación de la variable objetivo, *Tumor type*, a un formato numérico. Este paso es crucial por varias razones:

- **Compatibilidad**: Convertir la variable objetivo a formato numérico asegura su compatibilidad con una amplia gama de algoritmos de machine learning. La mayoría de estos algoritmos están diseñados para trabajar con datos numéricos, por lo que esta conversión es esencial para el correcto funcionamiento del modelo.
- **Eficiencia**: Los algoritmos de aprendizaje automático suelen funcionar más eficientemente con variables numéricas. Las operaciones matemáticas subyacentes en estos algoritmos son más rápidas y precisas cuando se utilizan números en lugar de cadenas de texto.
- **Precisión**: La conversión a formato numérico evita posibles errores de procesamiento que pueden ocurrir con variables categóricas en formato string. Al trabajar con números, se minimiza el riesgo de errores de codificación y se mejora la consistencia de los resultados.

Este enfoque de preprocesamiento no solo mejora la compatibilidad y eficiencia del modelo, sino que también asegura la precisión en la predicción de los distintos tipos de tumores.

Seguidamente, desarrollamos un código de transformación y combinación de datos con el objetivo de crear un nuevo DataFrame, que utilizaremos posteriormente para calcular la probabilidad total de predicción de los distintos tipos de tumores.

Este código realiza una serie de operaciones de preprocesamiento sobre un conjunto de datos de entrenamiento y prueba, utilizando librerías de Python como `numpy`, `pandas` y `sklearn`. El proceso incluye:

1. **Transformación y Combinación de Datos**: Utilizamos la función `combine_transformed_data_full` para combinar las características transformadas y la variable objetivo en un único DataFrame. Este enfoque nos permite trabajar con un conjunto de datos unificado que facilita las etapas posteriores de análisis y modelado.
2. **Preprocesamiento de Características**:
    - **Estandarización de Características Numéricas**: Las características numéricas se estandarizan para ajustar sus valores a una distribución con media cero y desviación estándar uno, lo cual es crucial para mejorar el rendimiento de muchos algoritmos de machine learning.
    - **Codificación de Características Categóricas**: Las características categóricas se transforman mediante técnicas como la codificación ordinal y la codificación one-hot, asegurando que estas variables puedan ser interpretadas adecuadamente por los modelos de machine learning.
3. **Transformación de la Variable Objetivo**: La variable objetivo, que representa el tipo de tumor, se convierte a valores numéricos utilizando `LabelEncoder`. Esto es fundamental por varias razones:
    - **Compatibilidad**: Asegura que la variable objetivo sea compatible con una amplia gama de algoritmos de machine learning.
    - **Eficiencia**: Los algoritmos de aprendizaje suelen funcionar más eficientemente con variables numéricas.
    - **Precisión**: Evita posibles errores de procesamiento que pueden ocurrir con variables categóricas en formato string.

Finalmente, el DataFrame combinado se guarda en un archivo Excel y se muestra una vista previa del mismo. Este proceso garantiza que los datos estén en el formato adecuado para el análisis posterior y la construcción de modelos predictivos, facilitando la predicción precisa de los distintos tipos de tumores.

	AFP (pg/ml)	Angiopoietin-2 (pg/ml)	AXL (pg/ml)	CA-125 (U/ml)	CA 15-3 (U/ml)	CA19-9 (U/ml)	CD44 (ng/ml)	CEA (pg/ml)	CYFRA 21-1 (pg/ml)	DKK1 (ng/ml)	...	sHER2/sEGFR2/sErbB2 (pg/ml)	sPECAM-1 (pg/ml)	TGFa (pg/ml)	Thrombospondin-2 (pg/ml)	TIMP-1 (pg/ml)	TIMP-2 (pg/ml)	Omega score	AJCC Stage	Sex_Male	Tumor type
0	-0.162730	-0.399967	0.161177	-0.155628	-0.208773	-0.126586	0.100087	-0.177265	-0.118556	-0.748740	...	0.613042	-0.739088	-0.190133	-0.501581	-0.709657	0.506555	-0.168524	0	1	1
1	-0.161972	-0.412345	-0.862979	-0.155856	-0.139035	-0.126858	-0.260427	-0.162648	-0.117902	-0.268988	...	-0.502477	-0.480916	-0.191864	-0.504778	-0.669592	-0.306917	-0.239250	1	1	4
2	-0.155496	-0.507819	-0.284533	-0.155194	-0.255232	0.053126	-0.630079	-0.184200	-0.115457	0.867268	...	-0.710243	-0.461172	-0.182823	-0.367130	-0.404068	-1.120939	-0.243808	1	1	1
3	-0.150629	-0.584907	-0.928337	-0.155787	-0.230770	-0.106049	0.047755	-0.181969	-0.011969	0.210765	...	-0.584788	-0.635555	-0.198597	-0.488231	-0.154741	1.505864	-0.042227	1	0	1
4	-0.147157	-0.464835	1.144186	-0.131803	0.317763	-0.085735	2.054673	-0.192859	-0.119066	-1.203243	...	0.827342	1.377542	-0.207830	-0.195666	-0.340628	1.733127	-0.215178	0	1	6

Fig. 7

A continuación, se presenta un análisis de la cantidad de datos disponibles para cada tipo de tumor. La siguiente tabla resume la distribución de los diferentes tipos de tumores en el conjunto de datos:

Tipo de Tumor	Cantidad
Colorectum	388
Breast	209
Lung	104
Pancreas	93
Stomach	68
Ovary	54
Esophagus	45
Liver	44

Fig. 8

La distribución de los datos, representada en la Fig. 8, revela un claro desbalance entre las distintas categorías de tumores. Observamos que ciertas categorías, como *Colorectum* y *Breast*, tienen una cantidad significativamente mayor de ejemplos en comparación con otras, como *Esophagus* y *Liver*. Este desbalance en los datos puede impactar negativamente el rendimiento de los modelos de machine learning, ya que tienden a sesgarse hacia las clases más representadas. Es crucial abordar este desbalance durante el preprocesamiento y la construcción de los modelos para asegurar una evaluación justa y precisa de todas las clases.

2.3.2. Evaluación de Modelos con Estrategias de Balanceo

En esta sección del proyecto, se procede a la evaluación y comparación del rendimiento de varios algoritmos de machine learning para la clasificación de tipos de tumores, utilizando técnicas de balanceo de datos para abordar conjuntos desbalanceados. El objetivo es seleccionar el modelo más adecuado que logre la mejor precisión y generalización posible.

Para implementar esta evaluación, se emplean bibliotecas como *scikit-learn* e *imblearn*, que proporcionan herramientas para la clasificación y técnicas de balanceo de *undersampling*. Se han definido funciones específicas para calcular métricas clave como precision, recall, F1-score y matriz de confusión, así como una puntuación global ponderada (*Global Score*) que facilita la comparación entre modelos. Cerramos esta fase del proyecto, guardando los datos obtenidos en un archivo formato excel, *model_results.xlsx* .

Los modelos de clasificación seleccionados para esta evaluación incluyen:

- **Logistic Regression**: Utilizado para problemas de clasificación binaria.
- **Decision Tree Classifier**: Ideal para clasificación multiclase con control de la profundidad del árbol.
- **Random Forest Classifier**: Implementa un bosque aleatorio para mejorar la precisión en clasificación multiclase.
- **K-Nearest Neighbors (KNN)**: Basado en vecinos más cercanos con métrica de distancia Manhattan.
- **AdaBoost Classifier**: Utiliza boosting para mejorar la precisión en clasificación multiclase.
- **Gradient Boosting Classifier**: Otra técnica de boosting para optimizar la clasificación multiclase.

Además de la selección de modelos, se incorporan estrategias de balanceo como parte del proceso de evaluación. Estas estrategias incluyen:

1. **RandomUnderSampler**

- **Propósito**: Esta técnica reduce el número de ejemplos en la clase mayoritaria seleccionando aleatoriamente muestras sin reemplazo para equilibrar el número de ejemplos en la clase minoritaria.
- **Ventaja**: Es sencillo y rápido de implementar, especialmente efectivo con conjuntos de datos grandes. Desventaja: Puede eliminar ejemplos importantes de la clase mayoritaria, lo que podría resultar en la pérdida de información relevante.
- **Tratamiento**: Undersampling.

2. **NearMiss**

- **Propósito**: NearMiss es una técnica de submuestreo que selecciona ejemplos de la clase mayoritaria basados en su proximidad a los ejemplos de la clase minoritaria.
- **Ventaja**: Conserva ejemplos representativos de la clase mayoritaria cerca de la frontera de decisión del clasificador.
- **Desventaja**: Similar al RandomUnderSampler, puede eliminar ejemplos significativos y puede ser costoso computacionalmente en conjuntos de datos grandes.
- **Tratamiento**: Undersampling.

3. **CondensedNearestNeighbour (CNN)**

- **Propósito**: CNN elimina ejemplos redundantes de la clase mayoritaria mientras conserva un subconjunto que representa bien la frontera de decisión del conjunto de datos.
- **Ventaja**: Reduce el tamaño del conjunto de datos sin perder ejemplos críticos para la clasificación.
- **Desventaja**: Puede ser computacionalmente costoso, especialmente con conjuntos de datos de alta dimensionalidad, y su eficacia puede variar.
- **Tratamiento**: Undersampling.

Estas estrategias de submuestreo (undersampling) se centran en reducir el número de muestras de la clase mayoritaria para equilibrar el conjunto de datos, mejorando así el rendimiento de los modelos al mitigar el sesgo hacia las clases dominantes.

```
Modelo: Logistic Regression
Cross-validated accuracy: 0.6182
Model performance for Training - Logistic Regression
- Accuracy: 0.7212
- F1 score: 0.7054
- Precision: 0.7429
- Recall: 0.7212
Confusion Matrix:
 [[118  22   0   0   2   1   2   1]
 [ 20 241   0   0   5   0   3   2]
 [  1  19   7   0   3   0   0   1]
 [  2   9   1  17   1   0   0   1]
 [  7  31   0   0  35   0   0   0]
 [  1  12   0   0   0  25   0   0]
 [  3  12   0   0   0   0  50   0]
 [  2  23   1   0   8   0   0  14]]
Model performance for Test - Logistic Regression
- Accuracy: 0.6722
- F1 score: 0.6513
- Precision: 0.6830
- Recall: 0.6722
Confusion Matrix:
 [[47 12  0  0  3  0  1  0]
 [12 99  0  0  2  1  0  3]
 [ 1  7  4  0  0  0  1  1]
 [ 2  6  0  5  0  0  0  0]
 [ 3 13  0  0 15  0  0  0]
 [ 0  5  0  1  0 10  0  0]
 [ 0  6  0  0  1  0 21  0]
 [ 2 13  0  1  1  0  1  2]]

Modelo: Decision Tree
```

Fig. 9

2.3.3. Interpretación de modelos

Cargamos el archivo *dataframe model_results* que generamos anteriormente y detectamos los 3 mejores modelos en la fase de prueba. Para ello ordenamos usando la puntuación obtenida en el parámetro *Global Score*:

	Model	Set	Accuracy	Precision	Recall	F1-Score	Global Score
1	Logistic Regression	Test	0.672185	0.683021	0.672185	0.651319	66.97
11	Gradient Boosting	Test	0.639073	0.619969	0.639073	0.615271	62.83
7	KNN	Test	0.609272	0.632833	0.609272	0.585354	60.92

Fig. 10

2.3.4. Exploración de Variables Relevantes en la Clasificación de Tipos de Tumores

Para identificar las variables más influyentes en la clasificación de tipos de tumores, se empleó un modelo *Random Forest* sobre un conjunto de datos preprocesado. Inicialmente, se aseguró la integridad de los datos eliminando cualquier instancia con valores faltantes del conjunto combinado y transformado.

Luego, se procedió a separar las características (variables predictoras) de la etiqueta de destino, que en este caso es el tipo de tumor. El conjunto de datos se dividió en datos de entrenamiento y prueba, utilizando una proporción del 80% para entrenamiento y 20% para prueba.

El modelo *Random Forest* se configuró con 250 árboles y se ajustó a los datos de entrenamiento. Posteriormente, se calculó la importancia de cada característica utilizando el atributo `feature_importances_` del modelo. Estas importancias se ordenaron de manera descendente y se visualizaron mediante un gráfico de barras utilizando la biblioteca *seaborn*, facilitando la comprensión de la contribución relativa de cada variable en la clasificación.

Finalmente, los resultados se almacenaron en un archivo CSV llamado *feature_importances.csv*, proporcionando una referencia para análisis posteriores y facilitando la comunicación de los hallazgos. Este enfoque no solo permite identificar las variables más relevantes en la clasificación de tumores, sino que también mejora la interpretación y visualización de los resultados mediante gráficos claros y detallados.

A continuación se muestra la tabla que presenta las 20 características más importantes según el modelo Random Forest:

	Feature	Importance
31	sFas (pg/ml)	0.044544
33	sHER2/sEGFR2/sErbB2 (pg/ml)	0.038370
4	CA 15-3 (U/ml)	0.034747
5	CA19-9 (U/ml)	0.033765
3	CA-125 (U/ml)	0.033653
38	TIMP-2 (pg/ml)	0.032779
35	TGFa (pg/ml)	0.032361
41	Sex_Male	0.032141
21	Leptin (pg/ml)	0.031827
19	IL-8 (pg/ml)	0.031686

	Feature	Importance
18	IL-6 (pg/ml)	0.028046
0	AFP (pg/ml)	0.027086
15	GDF15 (ng/ml)	0.026415
29	Prolactin (pg/ml)	0.026332
17	HGF (pg/ml)	0.025399
6	CD44 (ng/ml)	0.025003
23	Midkine (pg/ml)	0.024707
36	Thrombospondin-2 (pg/ml)	0.024402
37	TIMP-1 (pg/ml)	0.023657
16	HE4 (pg/ml)	0.023151

Fig. 11

Este enfoque metodológico proporciona una manera efectiva de identificar y visualizar qué variables son más relevantes para la clasificación de tumores, promoviendo una mejor comprensión y análisis de los datos biomédicos.

2.3.5. Entrenamiento y Evaluación de Modelos XGBoost y LightGBM para la Detección de Tipos de Cáncer con Objetivo Desbalanceado

XGBoost (Extreme Gradient Boosting) y *LightGBM* (Light Gradient Boosting Machine) son poderosos algoritmos de boosting ampliamente utilizados en la detección de tipos de cáncer, incluso en conjuntos de datos con desequilibrio de clases. Aquí se detallan las características que hacen que estos modelos sean efectivos:

XGBoost

1. **Boosting por Gradiente**:

XGBoost utiliza un proceso secuencial de construcción de árboles, donde cada árbol corrige los errores del anterior, optimizando así la capacidad predictiva.

2. **Manejo de Datos Desbalanceados**:

Incorpora parámetros como *scale_pos_weight* para ajustar el peso de las clases y abordar el desbalance de datos, lo cual es crucial en la detección de cáncer donde las clases pueden estar desproporcionadamente representadas.

3. **Regularización**:

Implementa regularización L1 y L2 para evitar el sobreajuste, mejorando la generalización del modelo en nuevos datos.

4. **Eficiencia Computacional**:

Optimizado para operaciones eficientes, aprovechando características de hardware y software para manejar grandes volúmenes de datos con rapidez.

5. **Importancia de Características**:

Proporciona medidas de importancia de características que ayudan a identificar biomarcadores relevantes en la clasificación de tipos de cáncer.

6. **Rendimiento y Análisis del Rendimiento del Modelo XGBoost**:

El rendimiento del modelo XGBoost para la detección de tipos de cáncer en un conjunto de datos desbalanceado muestra resultados prometedores. A continuación, se presenta un análisis detallado de las métricas obtenidas:

- **Accuracy**: 0.6468

La métrica accuracy del modelo XGBoost es de 0.6468, lo que indica que aproximadamente el 64.68% de las predicciones fueron correctas. Esta métrica proporciona una visión general de la capacidad del modelo para clasificar correctamente los casos. Aunque es una métrica útil, puede no reflejar completamente el rendimiento del modelo en un contexto de datos desbalanceados.

- **F1 score**: 0.6174

El F1 score es de 0.6174, lo que refleja un equilibrio entre la precision y el recall. Este valor sugiere que el modelo tiene un rendimiento razonable, considerando tanto su capacidad para identificar correctamente los casos positivos (recall) como la proporción de predicciones positivas correctas (precision). En escenarios de desbalance de clases, el F1 score es una métrica crucial, y un valor de 0.6174 indica un desempeño satisfactorio del modelo XGBoost.

- **Precision**: 0.6233

La métrica precision del modelo es de 0.6233, lo que significa que el 62.33% de las predicciones positivas del modelo fueron correctas. Una alta precisión es importante en la detección de cáncer para reducir los falsos positivos, evitando diagnósticos erróneos y tratamientos innecesarios. Un valor de 0.6233 indica que el modelo tiene una precisión moderada-alta, lo que es positivo, aunque aún hay espacio para mejorar.

- **Recall**: 0.6468

El recall, también conocido como sensibilidad, es de 0.6468. Este valor indica que el modelo identificó correctamente el 64.68% de los casos positivos reales. En la detección de cáncer, un alto valor de recall es esencial para minimizar los falsos negativos, asegurando que la mayoría de los casos de cáncer se detecten. Un recall de 0.6468 sugiere que el modelo XGBoost tiene una buena capacidad para detectar casos positivos, aunque también indica que aún hay un 35.32% de casos positivos que no fueron detectados.

7. **Conclusión** El análisis del rendimiento del modelo XGBoost revela un desempeño bastante sólido en la detección de tipos de cáncer en un conjunto de datos desbalanceado. Con una precisión del 64.68%, un F1 score de 0.6174, una precisión de 0.6233 y un recall de 0.6468, el modelo muestra que puede diferenciar entre las clases con un nivel de exactitud razonable. Estos resultados indican que XGBoost es eficaz para esta tarea, aunque se pueden realizar mejoras adicionales para aumentar su rendimiento. Técnicas como el ajuste de hiperparámetros, el uso de métodos de preprocesamiento avanzados y la combinación con otros enfoques de machine learning podrían ayudar a mejorar aún más estos resultados. Este análisis sugiere que XGBoost es una herramienta valiosa para la detección de tipos de cáncer, proporcionando una base sólida para trabajos futuros.

LightGBM

1. **Boosting por Hojas**:

LightGBM adopta un enfoque basado en hojas en lugar de niveles, lo que le permite manejar mejor los datos desbalanceados al centrarse en las hojas con mayores errores.

2. **Tratamiento del Desbalanceo**:

Al igual que XGBoost, ofrece opciones como *is_unbalance* y *scale_pos_weight* para ajustar automáticamente el modelo y tratar eficazmente el desbalance de clases.

3. **Velocidad y Eficiencia**:

Conocido por su rapidez y eficiencia en conjuntos de datos grandes, utiliza métodos como el aprendizaje basado en histogramas para acelerar el entrenamiento.

4. **Escalabilidad**:

Capaz de manejar conjuntos de datos de alta dimensionalidad y grandes volúmenes gracias a su diseño optimizado para memoria y capacidad de paralelización.

5. **Reducción del Overfitting**:

Incorpora estrategias avanzadas de regularización y ajuste de hiperparámetros para mitigar el sobreajuste, mejorando así la precisión en datos de prueba.

6. **Rendimiento y Análisis del Rendimiento del Modelo LightGBM**:

El rendimiento del modelo LightGBM para la detección de tipos de cáncer en un conjunto de datos desbalanceado muestra resultados que, aunque no sobresalientes, proporcionan una base sólida para futuras mejoras. A continuación, se presenta un análisis detallado de las métricas obtenidas:

- **Accuracy**: 0.5821

La métrica accuracy del modelo LightGBM es de 0.5821, lo que indica que aproximadamente el 58.21% de las predicciones fueron correctas. En un contexto de datos desbalanceados, esta métrica puede ser engañosa, ya que puede estar sesgada por la clase mayoritaria. Sin embargo, proporciona una visión general inicial de la capacidad del modelo para clasificar correctamente los casos.

- **F1 score**: 0.5029

El F1 score es de 0.5029, lo que refleja un equilibrio entre la precision y el recall. Este valor indica que el modelo tiene un rendimiento moderado al considerar tanto la capacidad de identificar correctamente los casos positivos (recall) como la proporción de predicciones positivas correctas (precision). Dado que el F1 score es útil en escenarios con clases desbalanceadas, su valor sugiere que el modelo LightGBM tiene un rendimiento aceptable, aunque hay margen para mejoras significativas.

- **Precision**: 0.5293

La métrica precision del modelo es de 0.5293, lo que significa que el 52.93% de las predicciones positivas del modelo fueron correctas. Esta métrica es crucial en la detección de cáncer, ya que un alto valor de precisión reduce la cantidad de falsos positivos, evitando diagnósticos incorrectos que pueden llevar a tratamientos innecesarios. Un valor de 0.5293 indica que el modelo tiene una precisión moderada, lo que sugiere la necesidad de afinar el modelo para reducir los falsos positivos.

- **Recall**: 0.5821

El recall, también conocido como sensibilidad, es de 0.5821. Este valor indica que el modelo identificó correctamente el 58.21% de los casos positivos reales. En la detección de cáncer, un alto valor de recall es esencial para minimizar los falsos negativos, asegurando que la mayoría de los casos de cáncer se detecten. Aunque un recall de 0.5821 es razonable, indica que hay un 41.79% de casos positivos que el modelo no detectó, lo que es un área crítica a mejorar para asegurar diagnósticos más completos.

7. **Conclusión** El análisis del rendimiento del modelo LightGBM revela un desempeño moderado en la detección de tipos de cáncer en un conjunto de datos desbalanceado. Con un accuracy del 58.21%, un F1 score de 0.5029, un precision de 0.5293 y un recall de 0.5821, el modelo muestra que puede diferenciar entre las clases, aunque no con una alta exactitud. Para mejorar su efectividad, se deben considerar técnicas adicionales de preprocesamiento de datos, ajuste de hiperparámetros y potencialmente la combinación con otros modelos o técnicas de ensemble. Este análisis sugiere que, aunque LightGBM tiene un buen punto de partida, hay oportunidades significativas para mejorar su rendimiento en futuros trabajos.

	Model	Set	Accuracy	Precision	Recall	F1-Score	Global Score
0	XGBoost	Training	0.822139	0.832805	0.822139	0.813388	82.26
1	XGBoost	Test	0.646766	0.623312	0.646766	0.617399	63.36
2	LightGBM	Training	0.708955	0.799136	0.708955	0.657890	71.87
3	LightGBM	Test	0.582090	0.529273	0.582090	0.502930	54.91

Fig. 12

Estos resultados detallan la precisión, recall, F1-score y la puntuación global para los conjuntos de entrenamiento y prueba de los modelos XGBoost y LightGBM utilizados en la detección de tipos de cáncer. La misma tabla está disponible en el archivo CSV *XGBoost_LightGBM.csv* para análisis adicional.

2.4. Ensamble de modelos (VotingClassifier)

Hemos decidido recurrir a un Ensamble de Modelos para mejorar la detección de tipos de cáncer.

VotingClasssiffier es una técnica poderosa para mejorar el rendimiento y la robustez de los modelos de machine learning. En el contexto de detección de tipos de cáncer con variables objetivo desbalanceadas, el ensamble puede proporcionar varias ventajas significativas:

1. **Reducción del Overfitting**

- **Promedio de Errores**: Al combinar varios modelos, los errores específicos de cada modelo tienden a cancelarse entre sí. Esto ayuda a reducir el overfitting, ya que los modelos individuales pueden sobreajustarse a ruidos o patrones específicos del conjunto de entrenamiento, pero estos errores se compensan cuando se utilizan múltiples modelos.

2. **Mejora de la Generalización**

- **Diversidad de Modelos**: Diferentes modelos pueden capturar diferentes aspectos de los datos. Por ejemplo, *XGBoost* y *LightGBM*, aunque ambos son métodos de boosting, tienen diferentes mecanismos internos que pueden captar distintas características del conjunto de datos. Un ensamble puede aprovechar estas diferencias y mejorar la capacidad de generalización del modelo final.

3. *Estabilidad y Robustez*

- **Promedio de Resultados**: La combinación de múltiples modelos tiende a producir resultados más estables y robustos frente a variaciones en los datos. Esto es especialmente importante en aplicaciones críticas como la detección de cáncer, donde las predicciones erróneas pueden tener consecuencias graves.

4. *Manejo de Datos Desbalanceados*

- **Balance de Clases**: Al utilizar técnicas de ensamble, se pueden diseñar estrategias específicas para manejar datos desbalanceados, como ajustar los pesos de las clases o utilizar técnicas de resampling dentro del ensamble. Esto puede mejorar la sensibilidad y especificidad de las predicciones para la clase minoritaria.

2.4.1. ¿Por que hemos escogido los siguientes modelos?

A continuación reflejamos dichas razones para usar estos modelos:

- **Mejores Métricas**: Estas combinaciones han demostrado ofrecer las mejores métricas de rendimiento en nuestras pruebas, lo que indica que son capaces de manejar bien los datos desbalanceados y proporcionar predicciones precisas.
- **Diversidad en Modelos de Boosting**: Cada uno de estos algoritmos tiene mecanismos internos ligeramente diferentes y fortalezas únicas que, cuando se combinan, pueden mejorar la capacidad de generalización del modelo final.
- **Reducción del Overfitting**: Al combinar varios modelos, se pueden cancelar los errores específicos de cada uno, reduciendo el riesgo de sobreajuste y mejorando la robustez del modelo.
- **Robustez y Estabilidad**: La combinación de múltiples modelos tiende a producir resultados más estables y robustos frente a variaciones en los datos, lo cual es crucial en aplicaciones críticas como la detección de cáncer.

En resumen, la elección de estas combinaciones de modelos de ensamble está respaldada por su rendimiento superior en términos de métricas y su capacidad para manejar datos desbalanceados, lo que los hace ideales para la tarea de detección de tipos de cáncer en nuestro caso específico.

2.4.2. Explicación del Pipeline de VotingClassiffier

Este pipeline de machine learning está diseñado para entrenar y evaluar varios modelos combinados de clasificación para predecir el tipo de tumor basado en un conjunto de características importantes. A continuación se presenta una explicación detallada del propósito y funcionamiento de cada parte del código:

1. **Importaciones de Librerías**

Se utilizan diversas librerías, como *pandas* para manejar y manipular datos en forma de DataFrame, y *sklearn* para dividir los datos en conjuntos de entrenamiento y prueba, así como para utilizar los modelos *Gradient Boosting* y *VotingClassifier*. También se emplean *lightgbm* y *xgboost* para usar los modelos *LightGBM* y *XGBoost*, y *sklearn.metrics* para evaluar el rendimiento de los modelos con métricas como *accuracy, precision, recall y F1-score*.

2. **Inicialización del DataFrame para Resultados**

Se inicializa un DataFrame vacío llamado `tabla_results_df` para almacenar los resultados de las evaluaciones de los modelos.

3. **Preprocesamiento de Datos**

Para asegurar que no haya datos faltantes, se eliminan las filas con datos faltantes del DataFrame utilizando `dropna()`.

4. **Selección de Características Importantes**

Se seleccionan las características más relevantes para entrenar el modelo, incluyendo varias mediciones de proteínas y marcadores tumorales.

5. **Separación de Características y Etiquetas**

Las características seleccionadas se almacenan en una variable `X`, y las etiquetas del tipo de tumor se almacenan en una variable `y`.

6. **División en Conjuntos de Entrenamiento y Prueba**

Se divide el conjunto de datos en un 80% para entrenamiento y un 20% para prueba, manteniendo la distribución de clases mediante `stratify`.

7. Definición de Modelos Base

Se definen tres modelos base: *Gradient Boosting*, *LightGBM* y *XGBoost*, cada uno con parámetros específicos.

8. **Definición de Combinaciones de Modelos**

Se crean combinaciones de los modelos base para usar en el *VotingClassifier*, que combina varios modelos mediante votación suave (`voting='soft'`).

9. **Entrenamiento y Evaluación de Cada Combinación**

Para cada combinación de modelos, se crea un *VotingClassifier* y se entrena y evalúa utilizando una función que entrena el modelo combinado y evalúa su rendimiento en los conjuntos de entrenamiento y prueba. Los resultados se almacenan en `tabla_results_df`.

10. **Guardado de Resultados**

Finalmente, los resultados almacenados en `tabla_results_df` se guardan en un archivo CSV llamado '*Voting_classifier.csv*'.

Este pipeline permite evaluar la efectividad de combinaciones de modelos de clasificación avanzados (*Gradient Boosting*, *LightGBM* y *XGBoost*) para predecir el tipo de tumor, utilizando un conjunto específico de características importantes. La evaluación incluye métricas como *accuracy, precision, recall y F1-score*, y los resultados se almacenan y exportan para su posterior análisis.

	Model	Set	Accuracy	Precision	Recall	F1-Score	Global Score
0	Gradient Boosting + LightGBM	Training	1.000000	1.000000	1.000000	1.000000	100.00
1	Gradient Boosting + LightGBM	Test	0.726368	0.699776	0.726368	0.703220	71.39
2	Gradient Boosting + XGBoost	Training	1.000000	1.000000	1.000000	1.000000	100.00
3	Gradient Boosting + XGBoost	Test	0.696517	0.681829	0.696517	0.679164	68.85
4	LightGBM + XGBoost	Training	1.000000	1.000000	1.000000	1.000000	100.00
5	LightGBM + XGBoost	Test	0.706468	0.691023	0.706468	0.685159	69.73
6	Gradient Boosting + LightGBM + XGBoost	Training	1.000000	1.000000	1.000000	1.000000	100.00
7	Gradient Boosting + LightGBM + XGBoost	Test	0.721393	0.704977	0.721393	0.701901	71.24

Fig. 13

2.4.3. Análisis de Resultados del Ensamble de Modelos (VotingClassifier)

En este análisis, examinamos los resultados obtenidos de diferentes combinaciones de modelos en el archivo `Voting_classifier.csv`. Se evaluaron cuatro combinaciones de modelos: *Gradient Boosting + LightGBM*, *Gradient Boosting + XGBoost*, *LightGBM + XGBoost*, y *Gradient Boosting + LightGBM + XGBoost*. Los resultados se presentan para los conjuntos de datos de entrenamiento y prueba.

1. **Vista Previa de los Resultados**

Los datos se estructuran en un DataFrame que contiene las siguientes columnas:

- `Model`: Combinación de modelos utilizada.
- `Set`: Conjunto de datos (entrenamiento o prueba).
- `Accuracy`: Precisión de las predicciones.
- `Precision`: Precisión de las predicciones.
- `Recall`: Sensibilidad de las predicciones.
- `F1-Score`: Puntaje F1, que combina precisión y recall.
- `Global Score`: Puntaje global de la combinación de modelos.

2. **Rendimiento en el Conjunto de Entrenamiento**

Para cada combinación de modelos, el rendimiento en el conjunto de entrenamiento es perfecto, con todas las métricas (accuracy, precision, recall, F1-Score y global score) alcanzando el valor máximo posible:

- **Gradient Boosting + LightGBM**: Todas las métricas son 1.000.
- **Gradient Boosting + XGBoost**: Todas las métricas son 1.000.
- **LightGBM + XGBoost**: Todas las métricas son 1.000.
- **Gradient Boosting + LightGBM + XGBoost**: Todas las métricas son 1.000.

Este resultado sugiere un sobreajuste (overfitting), ya que el modelo se ajusta perfectamente a los datos de entrenamiento.

3. **Rendimiento en el Conjunto de Prueba**

Al evaluar los modelos en el conjunto de prueba, observamos una disminución en las métricas, indicando que los modelos no generalizan tan bien como en el conjunto de entrenamiento:

- **Gradient Boosting + LightGBM**:
  - Accuracy: 0.726
  - Precision: 0.700
  - Recall: 0.726
  - F1-Score: 0.703
  - Global Score: 71.39
- **Gradient Boosting + XGBoost**:
  - Accuracy: 0.697
  - Precision: 0.682
  - Recall: 0.697
  - F1-Score: 0.679
  - Global Score: 68.85
- **LightGBM + XGBoost**:
  - Accuracy: 0.706
  - Precision: 0.691
  - Recall: 0.706
  - F1-Score: 0.685
  - Global Score: 69.73
- **Gradient Boosting + LightGBM + XGBoost**:
  - Accuracy: 0.721
  - Precision: 0.705
  - Recall: 0.721
  - F1-Score: 0.702
  - Global Score: 71.24

4. **Comparación de Modelos**

Al comparar las diferentes combinaciones de modelos, observamos que la combinación de *Gradient Boosting + LightGBM* obtiene las mejores métricas en el conjunto de prueba, seguida de cerca por *Gradient Boosting + LightGBM + XGBoost*. La combinación *Gradient Boosting + XGBoost* tiene el rendimiento más bajo.

5. **Interpretación de Resultados**

- **Reducción del Overfitting**: Aunque todos los modelos muestran un excelente rendimiento en el conjunto de entrenamiento, su desempeño disminuye en el conjunto de prueba. Esto indica que los modelos podrían estar sobreajustando los datos de entrenamiento.
- **Mejora de la Generalización**: La combinación de *Gradient Boosting + LightGBM* parece generalizar mejor a los datos no vistos, obteniendo las métricas más altas en el conjunto de prueba.
- **Diversidad de Modelos**: Las diferencias en las métricas de rendimiento sugieren que las combinaciones de modelos capturan diferentes aspectos de los datos. *LightGBM* y *XGBoost* en conjunto no parecen mejorar significativamente el rendimiento en comparación con las otras combinaciones.
- **Robustez y Estabilidad**: La combinación de tres modelos (*Gradient Boosting + LightGBM + XGBoost*) también muestra buenos resultados, lo que indica que agregar más diversidad a los modelos puede mejorar la estabilidad del ensamble.

6. **Conclusiones**

- La combinación de *Gradient Boosting + LightGBM* ofrece el mejor rendimiento en términos de *accuracy, precision, recall y F1-Score* en el conjunto de prueba.
- Es necesario abordar el sobreajuste observando técnicas de regularización o utilizando un conjunto de validación cruzada más robusto.
- Futuras mejoras pueden incluir el ajuste de hiperparámetros y la exploración de otras técnicas de ensamble o modelos base adicionales.

Este análisis nos proporciona una comprensión clara del rendimiento actual de los modelos y sugiere direcciones para futuras mejoras en la detección de tipos de cáncer utilizando ensambles de modelos.

2.5. Tratamiento del overfitting

Tal y como hemos observado en el apartado anterior, nos hemos encontrado con el problema del overfitting, por lo que hemos actuado en consecuencia redactando una solución para ello.

2.5.1. Explicación del Código:

- **DataFrame *tabla_results_df***:

Se inicializa un DataFrame vacío con columnas específicas (*'Model', 'Set', 'Accuracy', 'Precision', 'Recall', 'F1-Score', 'Global Score'*) para almacenar los resultados de la evaluación de los modelos.

- **Preprocesamiento de Datos**:

Nos aseguramos de que no haya datos nulos eliminando las filas correspondientes del DataFrame *df_combined_transformed*. Se seleccionan las características importantes (*important_features*) que ya detectamos anteriormente en la búsqueda de las variables que aportan más información, para su próximo uso para el entrenamiento y la evaluación de los modelos.

- **División de Datos**:

Utilizando *train_test_split*, se dividen los datos en conjuntos de entrenamiento (*X_train, y_train*) y prueba (*X_test, y_test*). Esto es crucial para evaluar la capacidad de generalización de los modelos.

- *Definición de Modelos*:

Se definen tres modelos base: *GradientBoostingClassifier*, *LGBMClassifier* de *LightGBM* y *XGBClassifier* de *XGBoost*, cada uno con parámetros específicos para optimizar su rendimiento en el problema dado.

- **VotingClassifier**:

Se crea un *VotingClassifier* utilizando la técnica de voto suave (voting='soft'), que combina los tres modelos base (*gb, lgbm, xgb_clf*) para mejorar la precisión y robustez general del sistema predictivo.

- **Función *entrenar_y_evaluar_modelo***:

Esta función encapsula el proceso de entrenamiento y evaluación de un modelo dado. Toma como entrada los conjuntos de entrenamiento y prueba, el modelo a evaluar (*model*) y su nombre (*model_name*) para identificación en los resultados. Después de entrenar el modelo con *model.fit(X_train, y_train)*, realiza predicciones en *X_test* y calcula métricas clave como *precision, recall, F1-score y accuracy* utilizando funciones de *sklearn* (*accuracy_score, precision_score, recall_score, f1_score*).

Global Score se calcula promediando estas métricas para proporcionar una evaluación comprensiva del desempeño del modelo.

- *Almacenamiento de Resultados*:

Los resultados de la evaluación se añaden al DataFrame *tabla_results_df*, que posteriormente se guarda como un archivo CSV (*'Gradient_Boosting_LightGBM_XGBoost.csv'*). Esto facilita el análisis posterior y la comparación de modelos en términos de su rendimiento predictivo.

	Model	Set	Accuracy	Precision	Recall	F1-Score	Global Score
0	Gradient Boosting + LightGBM + XGBoost	Training	0.931592	0.934897	0.931592	0.929895	93.20
1	Gradient Boosting + LightGBM + XGBoost	Test	0.681592	0.664823	0.681592	0.643653	66.79

Fig. 14

El análisis del rendimiento del modelo "Gradient Boosting + LightGBM + XGBoost" se presenta en la tabla mostrada. Esta tabla resume las métricas clave obtenidas durante el entrenamiento y la evaluación del modelo en conjuntos de datos separados para entrenamiento y prueba.

2.5.2. Análisis de Resultados del Modelo

- **Conjunto de Entrenamiento**
- **Accuracy**: 93.16%
- **Precision**: 93.49%
- **Recall**: 93.16%
- **F1-Score**: 92.99%
- **Global Score**: 93.20

Estos resultados indican que el modelo logra una alta precisión y capacidad de generalización en los datos utilizados para entrenamiento, con métricas muy similares entre accuracy, precisin y recall. F1 Score, que combina precision y recall, es también muy robusto, reflejando un buen equilibrio entre ambas métricas.

- **Conjunto de Prueba**
- **Accuracy**: 68.16%
- **Precision**: 66.48%
- **Recall**: 68.16%
- **F1-Score**: 64.37%
- **Global Score**: 66.79

Comparando con el conjunto de entrenamiento, observamos una caída significativa en todas las métricas evaluadas en el conjunto de prueba. Esto sugiere que el modelo puede estar experimentando cierta dificultad para generalizar a datos no vistos durante el entrenamiento, lo cual es común en problemas complejos de aprendizaje automático.

- **Interpretación General**

Los resultados reflejan un buen desempeño del modelo en el conjunto de entrenamiento, donde logra altas métricas de precisión y capacidad predictiva. Sin embargo, en el conjunto de prueba, vemos una reducción en el rendimiento, indicando posibles áreas de mejora en términos de generalización y robustez del modelo.

Estos hallazgos son cruciales para entender las fortalezas y limitaciones del modelo *"Gradient Boosting + LightGBM + XGBoost"*, proporcionando una base sólida para iteraciones adicionales en la mejora del modelo y la exploración de estrategias para mitigar el overfitting y mejorar la capacidad de generalización en datos nuevos.

En conclusión, mientras que el modelo muestra promesas en datos de entrenamiento, es fundamental seguir refinándolo y validándolo con datos adicionales para garantizar su eficacia en aplicaciones del mundo real.

Este código realiza un análisis de clustering utilizando varios algoritmos y evalúa su desempeño utilizando métricas específicas. Aquí está el análisis paso a paso:

2.6. Modelos de Aprendizaje No supervisado

Como has ahora hemos trabajado con la modalidad de machine learning de aprendizaje supervisado, ahora vamos a tratar la faceta de aprendizaje no supervisado. A continuación describimos el código usado.

1. **Librerías Importadas**

- **sklearn.cluster**: Importa los algoritmos de clustering como *KMeans, DBSCAN, AgglomerativeClustering, Birch, MeanShift y OPTICS*.
- **sklearn.preprocessing**: Importa *StandardScaler* para estandarizar las características antes de aplicar clustering.
- **imblearn.under_sampling**: Importa técnicas de submuestreo como *RandomUnderSampler, NearMiss, y CondensedNearestNeighbour*.
- **imblearn.pipeline**: Importa *Pipeline* para construir un flujo de trabajo de aprendizaje automático que incluye preprocesamiento y modelado.
- **sklearn.metrics**: Importa métricas como *silhouette_score, homogeneity_score, completeness_score, y davies_bouldin_score* para evaluar la calidad del clustering.
- **pandas**: Importa para la manipulación de datos, incluyendo la creación y manipulación de DataFrames.
- **sklearn.model_selection**: Importa *train_test_split* para dividir los datos en conjuntos de entrenamiento y prueba.

2. **DataFrame y Función de Evaluación**

- **tabla_results_df**: DataFrame inicializado para almacenar los resultados de las métricas de clustering.
- **evaluar_clustering**: Función que calcula y muestra métricas de evaluación para clustering, como *silhouette_score, homogeneity_score, completeness_score, davies_bouldin_score*. Los resultados se agregan al DataFrame `tabla_results_df`.

3. **Definición de Modelos**

- **modelos**: Diccionario que contiene diferentes modelos de clustering con sus respectivos parámetros.
    - *KMeans* con 7 clusters.
    - *DBSCAN* con epsilon 0.3 y mínimo 5 muestras.
    - *Agglomerative Clustering* con 7 clusters.
    - *Birch* con 7 clusters.
    - *MeanShift* con un ancho de banda de 2.
    - *OPTICS* con mínimo 5 muestras.

4. **Preparación de Datos**

- **important_features**: Lista de características seleccionadas para el análisis de clustering.
- **X**: Conjunto de características seleccionadas del DataFrame `df_combined_transformed`.
- **y**: Etiquetas de tipo de tumor del DataFrame `df_combined_transformed`.

5. **Reducción de Dimensionalidad con PCA**

- **PCA**: Reducción de dimensionalidad a 2 componentes principales usando *PCA* (Análisis de Componentes Principales).

6. **Entrenamiento y Evaluación de Modelos**

- Itera sobre cada modelo en `modelos`:
    - Ajusta el modelo al conjunto de datos reducido por PCA (`X_pca`).
    - Obtiene las etiquetas de cluster resultantes (`labels`) y llama a `evaluar_clustering` para calcular y mostrar las métricas de evaluación para cada modelo.

7. **Guardar Resultados**

- **tabla_results_df.to_excel**: Guarda los resultados de las métricas de clustering en un archivo Excel llamado 'Results_No_supervisado.xlsx'.

	Model	Metric	Score
0	KMeans	Silhouette Score	0.507759
1	KMeans	Homogeneity	0.084649
2	KMeans	Completeness	0.161828
3	KMeans	Davies-Bouldin Index	0.470376
4	DBSCAN	Silhouette Score	0.321461
5	DBSCAN	Homogeneity	0.028120
6	DBSCAN	Completeness	0.153577
7	DBSCAN	Davies-Bouldin Index	1.738256
8	Agglomerative Clustering	Silhouette Score	0.519311
9	Agglomerative Clustering	Homogeneity	0.082623
10	Agglomerative Clustering	Completeness	0.146401
11	Agglomerative Clustering	Davies-Bouldin Index	0.520300
12	Birch	Silhouette Score	0.759317
13	Birch	Homogeneity	0.033259
14	Birch	Completeness	0.273319
15	Birch	Davies-Bouldin Index	0.235386
16	MeanShift	Silhouette Score	0.712432
17	MeanShift	Homogeneity	0.023990
18	MeanShift	Completeness	0.240052
19	MeanShift	Davies-Bouldin Index	0.226063
20	OPTICS	Silhouette Score	-0.191600

	Model	Metric	Score
21	OPTICS	Homogeneity	0.174898
22	OPTICS	Completeness	0.117642
23	OPTICS	Davies-Bouldin Index	1.910613

Fig. 15

2.6.1. Análisis de los Resultados del Clustering

El análisis de clustering evaluó varios algoritmos utilizando métricas específicas para entender cómo cada modelo agrupa los datos. A continuación se presentan los resultados clave:

1. **KMeans**:

- **Silhouette Score**: 0.5078
- **Homogeneidad**: 0.0846
- **Completeness**: 0.1618
- **Índice Davies-Bouldin**: 0.4704

El algoritmo KMeans muestra un Silhouette Score razonable y un Índice Davies-Bouldin bajo, indicando clusters bien definidos y compactos. Sin embargo, la homogeneidad y la completitud son relativamente bajas, sugiriendo que los clusters pueden no estar completamente separados o ser homogéneos en términos de etiquetas.

2. **DBSCAN**:

- **Silhouette Score**: 0.3215
- **Homogeneidad**: 0.0281
- **Completeness**: 0.1536
- **Índice Davies-Bouldin**: 1.7383

DBSCAN muestra un Silhouette Score inferior y un Índice Davies-Bouldin alto, lo que indica que puede haber clusters con diferentes densidades. La homogeneidad y la completitud también son bajas, lo que sugiere dificultades en la identificación de clusters homogéneos.

3. **Agglomerative Clustering**:

- **Silhouette Score**: 0.5193
- **Homogeneidad**: 0.0826
- **Completeness**: 0.1464
- **Índice Davies-Bouldin**: 0.5203

Este método muestra un buen Silhouette Score y un Índice Davies-Bouldin bajo, indicando clusters bien definidos y compactos. Sin embargo, al igual que KMeans, la homogeneidad y la completitud son relativamente bajas.

4. Birch:

- **Silhouette Score**: 0.7593
- **Homogeneidad**: 0.0333
- **Completeness**: 0.2733
- **Índice Davies-Bouldin**: 0.2354

Birch muestra el más alto Silhouette Score y un buen Índice Davies-Bouldin, indicando clusters bien definidos y compactos. La completitud es alta, pero la homogeneidad es baja, lo que sugiere que los clusters pueden no ser tan homogéneos en términos de etiquetas.

5. **MeanShift**:

- **Silhouette Score**: 0.7124
- **Homogeneidad**: 0.0240
- **Completeness**: 0.2401
- **Índice Davies-Bouldin**: 0.2261

MeanShift también muestra un buen Silhouette Score y un Índice Davies-Bouldin bajo, indicando clusters bien definidos y compactos. Sin embargo, al igual que otros métodos, la homogeneidad es baja.

6. **OPTICS**:

- **Silhouette Score**: -0.1916
- **Homogeneidad**: 0.1749
- **Completeness**: 0.1176
- **Índice Davies-Bouldin**: 1.9106

OPTICS muestra un Silhouette Score negativo, indicando una mala agrupación de datos. El alto Índice Davies-Bouldin también sugiere problemas en la separación y definición de clusters. La homogeneidad es relativamente alta, pero la completitud es baja.

2.6.2. Interpretación General

- **Silhouette Score**: Indica qué tan bien están separados los clusters. Valores cercanos a 1 son deseables.
- **Homogeneidad y Completeness**: Reflejan la uniformidad dentro de los clusters y la exhaustividad en la captura de todas las instancias de cada clase, respectivamente. Valores más altos son mejores.
- **Índice Davies-Bouldin**: Evalúa la separación efectiva entre los clusters. Valores más bajos indican una mejor separación.

En resumen, Birch muestra el mejor desempeño en términos de Silhouette Score y Índice Davies-Bouldin. Sin embargo, otros métodos exhiben fortalezas y debilidades diversas. La elección del algoritmo de clustering debe alinearse con los objetivos específicos del análisis y las características particulares del conjunto de datos.

- **Clusters No Bien Definidos**: Aunque Birch y MeanShift muestran resultados prometedores, la mayoría de los algoritmos no generan clusters de calidad suficiente para ofrecer una contribución significativa a nuestro problema.
- **Heterogeneidad de Clusters**: Las bajas homogeneidad y completitud sugieren una mezcla de clases dentro de los clusters, indicando que los modelos no capturan adecuadamente las estructuras subyacentes de los datos.

Notas

El análisis sugiere que el clustering no supervisado no aporta mucho valor a nuestro problema con este dataset. A pesar de que algunos algoritmos como Birch y MeanShift mostraron resultados decentes en ciertas métricas, la falta de clusters bien definidos y la baja homogeneidad y completitud indican que los modelos no están capturando estructuras significativas en los datos.

3. Plan para Mejorar el Dataset Utilizando UMAP y KMeans

Dado que los métodos de clustering no supervisado no han proporcionado métricas satisfactorias, hemos decidido adoptar una nueva estrategia para mejorar nuestro dataset. Nuestro objetivo principal es aumentar y enriquecer nuestro dataset, que actualmente cuenta con solo 1005 filas. Para ello, utilizaremos UMAP para la reducción de dimensionalidad y KMeans para identificar subgrupos, que luego añadiremos a nuestro dataset.

3.1. Objetivos y Pasos

- **Objetivo**

Mejorar el dataset actual de 1005 filas añadiendo información sobre subgrupos identificados mediante clustering.

- **Pasos**

1. **Reducción de Dimensionalidad con UMAP**

- **Objetivo**: Visualizar los datos en un espacio de menor dimensión para identificar patrones y estructuras subyacentes.
- **Método**: Aplicar *UMAP* (Uniform Manifold Approximation and Projection) para reducir la dimensionalidad de los datos a 2D o 3D.

2. **Identificación de Subgrupos con KMeans**

- **Objetivo**: Identificar subgrupos dentro de los datos reducidos dimensionalmente.
- **Método**: Aplicar *KMeans* en los datos reducidos por UMAP para identificar clusters o subgrupos.

3. **Añadir Subgrupos al Dataset Original**

- **Objetivo**: Enriquecer el dataset original añadiendo una nueva columna que indique el subgrupo al que pertenece cada punto.
- **Método**: Agregar los subgrupos identificados por KMeans como una nueva columna en el dataset original.

3.2. Beneficios de Esta Estrategia

- **Mayor Información**: Añadir subgrupos proporciona información adicional que puede ayudar a los modelos supervisados a capturar mejor las estructuras subyacentes en los datos.
- **Aumento de Datos**: Aunque no se aumentan las filas, se enriquece el dataset con nueva información, lo que puede mejorar el rendimiento de los modelos.
- **Mejor Visualización**: UMAP permite visualizar los datos en un espacio reducido, facilitando la identificación de patrones y anomalías.

3.3. Aplicar estrategia Reducción de Dimensionalidad con UMAP

En este análisis se emplea *UMAP* (Uniform Manifold Approximation and Projection), una técnica de reducción de dimensionalidad no lineal. El objetivo principal es visualizar datos complejos de características biológicas asociadas a diferentes tipos de cáncer en un espacio bidimensional.

El proceso comienza seleccionando las características numéricas relevantes del conjunto de datos, excluyendo cualquier columna no numérica. Estas características son esenciales para capturar las variaciones biológicas que diferencian los tipos de cáncer.

Posteriormente, se separan estas características de la variable objetivo, que en este caso es el tipo de tumor ('Tumor type'). Esta separación facilita la aplicación de UMAP, que se utiliza para reducir la alta dimensionalidad de los datos a solo dos dimensiones. Esta transformación preserva las estructuras locales de los datos, permitiendo una representación visual efectiva.

La visualización resultante se realiza mediante un gráfico de dispersión, donde cada punto representa una instancia de datos. Los puntos se colorean según el tipo de tumor para visualizar cómo se distribuyen los diferentes tipos en el espacio UMAP. Este enfoque visual ayuda a identificar agrupaciones naturales o patrones emergentes entre los tipos de cáncer basados en sus características biológicas.

La interpretación de los resultados se centra en la proximidad de los puntos en el gráfico de dispersión: puntos cercanos indican similitudes en las características biológicas entre los tipos de cáncer representados, mientras que puntos más separados reflejan diferencias más significativas. Esta visualización proporciona una perspectiva intuitiva y efectiva para explorar y comprender la estructura subyacente de los datos biológicos complejos.

En conclusión, UMAP facilita la exploración visual de datos complejos de cáncer, permitiendo a los investigadores identificar y comprender mejor las relaciones y diferencias entre diferentes tipos de cáncer basadas en sus características biológicas distintivas.

3.3.1. Análisis resultados

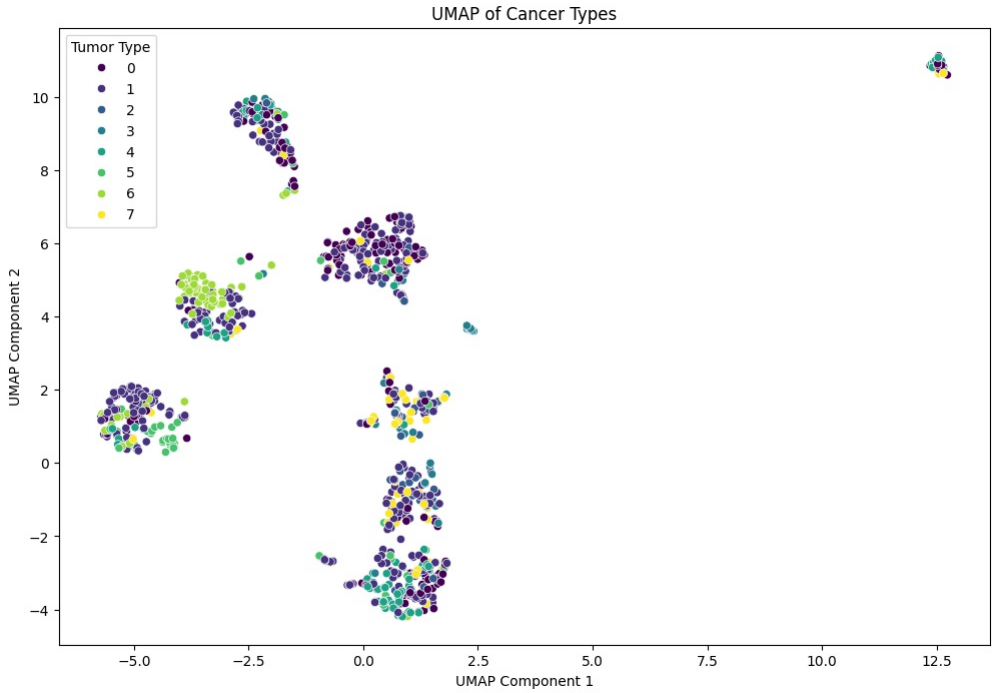


Fig. 16

Esta es la visualización de los resultados de la estrategia que aplicamos con la ayuda de UMAP. La interpretación de los resultados del clustering sugiere que pueden ser necesarios ajustes en los parámetros del algoritmo de clustering o en la selección de características para obtener resultados más distintivos y significativos.

Clustering con K-Means

Aplicamos K-Means junto a UMAP para identificar subgrupos en los datos, teniendo en cuenta los resultados anteriores.

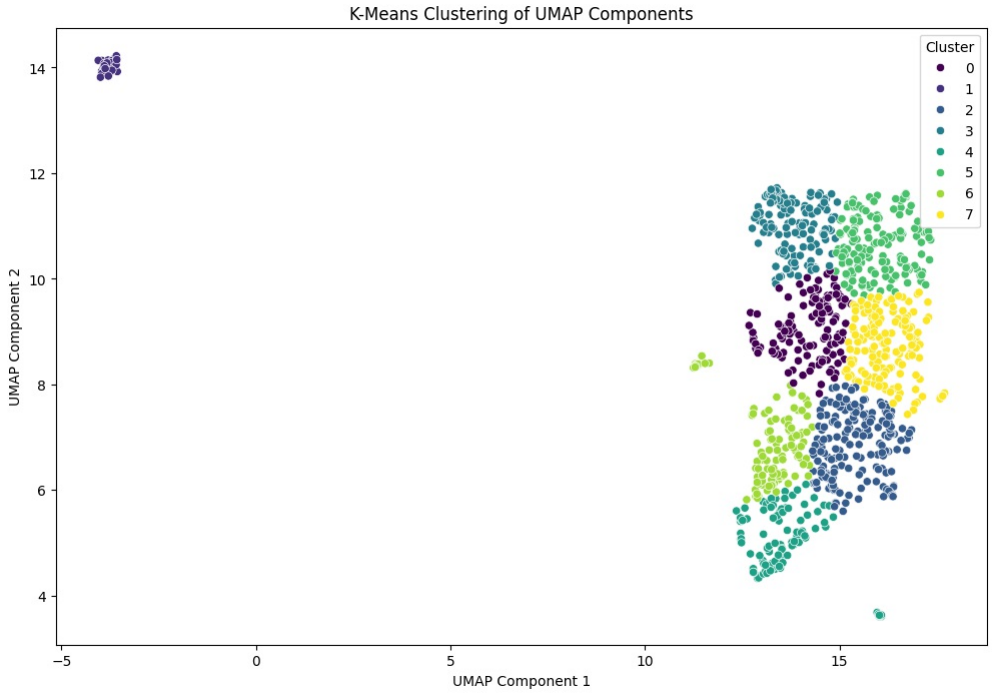


Fig. 17

Posibles Razones del Resultado obtenido

1. **Incoherencia en los Clusters**

- Es posible que los clusters generados no sean lo suficientemente informativos o consistentes con las etiquetas originales. Esto puede agregar ruido en lugar de valor al modelo.

2. **Complejidad del Modelo**

- La inclusión de clusters puede haber aumentado la complejidad del modelo sin proporcionar una ganancia de información significativa, lo que puede llevar a un sobreajuste o subajuste.

3.3.2. Explicación y Análisis de Ensemble de Modelos de Aprendizaje Automático para mejorar el rendimiento del clasificador

En este análisis, se emplea un ensemble de modelos de clasificación para predecir el tipo de tumor a partir de características transformadas y la inclusión de clusters como una nueva característica. A continuación se detallan los pasos clave del proceso:

1. **Preparación de Datos:**
    - Se importa el conjunto de datos utilizando `pandas`. Se añade una nueva característica llamada '*Cluster*', que representa los clusters obtenidos previamente, utilizando el algoritmo K-Means.
2. **Selección de Características y Variable Objetivo:**
    - Seleccionan las características relevantes para el modelo, excluyendo las que no contribuyen significativamente (`df_combined_transformed.columns[1:-2]`), y se añade '*Cluster*' como una característica adicional. La variable objetivo es '*Tumor type*'.
3. **División del Dataset:**
    - Se divide el dataset en conjuntos de entrenamiento y prueba utilizando `train_test_split()`. Esto permite evaluar la capacidad de generalización del modelo sobre datos no vistos.
4. **Definición de Modelos Individuales:**
    - Se definen varios modelos de clasificación, incluyendo `RandomForestClassifier`, `GradientBoostingClassifier`, `LGBMClassifier` de *LightGBM* y `XGBClassifier` de *XGBoost*. Cada modelo se inicializa con parámetros predeterminados y un estado aleatorio para asegurar la reproducibilidad de los resultados.
5. **Definición del Voting Classifier:**
    - Se crea un clasificador de votación (`VotingClassifier`) que combina los modelos individuales definidos anteriormente. En este caso, se utiliza el método de votación '*soft*', que considera las probabilidades predichas por cada modelo para tomar la decisión final.
6. **Entrenamiento del Voting Classifier:**
    - Se entrena el clasificador de votación utilizando los datos de entrenamiento (`X_train` y `y_train`).
7. **Evaluación del Modelo:**
    - Se realizan predicciones sobre el conjunto de prueba (`X_test`) utilizando VotingClassiffier entrenado. Se evalúa el rendimiento del modelo utilizando métricas como el `classification_report` y la matriz de confusión (`confusion_matrix`), que proporcionan información detallada sobre precision, recall y F1-score para cada clase de tumor.

Este enfoque de ensemble combina la fuerza predictiva de varios modelos individuales para mejorar la precisión y la capacidad de generalización del sistema de clasificación. La inclusión de clusters como una característica adicional permite explorar cómo la estructura de los datos agrupados puede influir en la precisión del modelo final. 8. **Análisis de Métricas**

Los resultados de la evaluación del modelo entrenado con LightGBM muestran un panorama mixto en términos de su desempeño para predecir diferentes clases de tumores. Aquí se detalla el análisis de los resultados basado en las métricas de evaluación y la matriz de confusión proporcionada:

1. **Precision:**
    - La precisión para cada clase varía significativamente:
        - Clase 0: 0.75
        - Clase 1: 0.70
        - Clase 2: 0.33
        - Clase 3: 1.00
        - Clase 4: 0.62
        - Clase 5: 0.90
        - Clase 6: 0.88
        - Clase 7: 0.38

    La precisión indica la proporción de predicciones positivas que fueron correctas respecto a todas las predicciones positivas realizadas por el modelo. Valores más altos indican una mejor capacidad del modelo para evitar falsos positivos.
2. **Recall:**
    - El recall para cada clase también muestra variaciones:
        - Clase 0: 0.79
        - Clase 1: 0.95
        - Clase 2: 0.11
        - Clase 3: 0.11
        - Clase 4: 0.38
        - Clase 5: 0.82
        - Clase 6: 0.79
        - Clase 7: 0.23

    El recall representa la proporción de verdaderos positivos que fueron correctamente identificados por el modelo respecto a todos los casos positivos reales. Valores más altos indican una mayor capacidad del modelo para detectar todos los casos positivos.
3. **F1-score:**
    - El F1-score, que es la media armónica de precision y recall, proporciona un balance entre ambas métricas:
        - Macro avg (promedio de todas las clases): 0.55
        - Weighted avg (ponderado por el soporte de cada clase): 0.68

    Un F1-score alto indica un buen equilibrio entre precision y recall, lo cual es deseable para un modelo de clasificación robusto.
4. **Support:**
    - El soporte indica el número de muestras reales que pertenecen a cada clase. Esto puede ser útil para interpretar la importancia relativa de cada clase en el conjunto de datos.
5. **Matriz de Confusión**

La matriz de confusión proporciona una visión más detallada de cómo el modelo clasifica las muestras en función de las clases reales:

```
[[33  4  0  0  4  0  1  0]
 [ 2 73  1  0  0  0  1  0]
 [ 0  5  1  0  0  0  0  3]
 [ 1  5  0  1  1  0  0  1]
 [ 6  6  0  0  8  0  0  1]
 [ 2  0  0  0  0  9  0  0]
 [ 0  3  0  0  0  1 15  0]
 [ 0  9  1  0  0  0  0  3]]
```

Cada fila de la matriz representa la clase real y cada columna representa la clase predicha por el modelo. Los números en la diagonal principal indican las predicciones correctas para cada clase. Observaciones importantes de la matriz de confusión incluyen:

- El modelo parece tener dificultades en predecir correctamente las clases minoritarias (por ejemplo, clases 2, 3, 5, y 7).
- Clases como la 1 y la 6 tienen una precisión y recall relativamente altos, indicando que el modelo es eficaz para estas clases específicas.
- Algunas confusiones significativas ocurren entre las clases 0 y 1, así como entre las clases 4 y 6, lo cual podría indicar cierta similitud en las características que el modelo tiene dificultades para distinguir.

Conclusión

En general, el modelo muestra un rendimiento promedio con un F1-score ponderado de 0.68. Aunque tiene una buena precisión y recall para algunas clases, también muestra áreas de mejora, especialmente en la capacidad para manejar clases minoritarias y para distinguir entre clases cercanas entre sí. Para mejorar el rendimiento, se podrían explorar estrategias como ajustes adicionales de hiperparámetros, consideración de pesos de clases o técnicas de resampling más sofisticadas.

Notas generales

El uso de un ensamble de modelos ha demostrado ser efectivo en mejorar el rendimiento del clasificador en nuestro dataset, incluso después de la incorporación de clusters como nuevas características. Este enfoque ha permitido obtener una mejor precisión y un equilibrio adecuado entre precisión y recall en la mayoría de las clases.

Sin embargo, dado que el rendimiento del modelo no ha mejorado significativamente, seguiremos explorando otras técnicas para mejorar nuestro dataset y el rendimiento del modelo.

3.4. Próximos Pasos

- **Optimización Adicional**
- Realizar una optimización más detallada de hiperparámetros para cada modelo individualmente y para el Voting Classifier en su conjunto. Esto se considerará más adelante, si el tiempo y la capacidad computacional lo permiten.
- **Manejo de Clases Minoritarias**
- Explorar técnicas adicionales para mejorar el rendimiento en las clases minoritarias. Esto incluye ajustar los pesos de las clases en los modelos y emplear técnicas avanzadas de resampling. Este será nuestro enfoque principal en los siguientes pasos.

4. Uso de CTGAN para Manejar Clases Minoritarias en un Dataset Pequeño

Debido a que nuestro dataset cuenta con solo 1005 filas, hemos decidido utilizar *Conditional Tabular Generative Adversarial Network (CTGAN)* para generar datos sintéticos y así abordar el problema de las clases minoritarias. Aquí explicamos en detalle lo que estamos haciendo, para qué sirven los GANs y por qué es importante en nuestro caso.

4.1. Objetivo

Generar datos sintéticos para las clases minoritarias en nuestro dataset para mejorar el balance de clases y proporcionar más datos para el entrenamiento de nuestros modelos supervisados.

4.2. Explicación de CTGAN y su Importancia

Qué son los CTGAN?

Son un tipo de red neuronal compuesta por dos modelos:

- **Generador**: Crea datos sintéticos a partir de ruido aleatorio.
- **Discriminador**: Distingue entre los datos reales y los datos sintéticos generados.

Estos dos modelos se entrenan de manera conjunta y competitiva: el generador intenta engañar al discriminador creando datos sintéticos realistas, mientras que el discriminador intenta mejorar su capacidad para diferenciar entre datos reales y sintéticos.

4.3. ¿Por qué usar CTGAN en nuestro Dataset?

1. **Incremento de Datos**: Nuestro dataset original es pequeño (1005 filas). CTGAN nos permiten generar datos sintéticos adicionales que pueden enriquecer nuestro dataset y mejorar el rendimiento de los modelos supervisados.
2. **Manejo de Clases Minoritarias**: Algunas clases en nuestro dataset están subrepresentadas.CTGAN pueden generar ejemplos sintéticos de estas clases minoritarias, ayudando a balancear el dataset y a mejorar la capacidad del modelo para aprender y generalizar sobre estas clases.
3. **Mejora del Modelo**: Al proporcionar más ejemplos para las clases minoritarias, esperamos mejorar la precisión y el recall de nuestro modelo en estas clases, reduciendo el sesgo y la varianza.

4.4. Generación de Datos Sintéticos y Análisis de resultados de CTGAN

Este código realiza varias tareas importantes relacionadas con la generación de datos sintéticos utilizando *CTGAN* y luego entrenando un Voting Classifier con modelos de *Gradient Boosting, LightGBM y XGBoost*.

1. **Preparación de Datos** Se carga y limpia el conjunto de datos para eliminar valores nulos y seleccionar características relevantes como `'sFas (pg/ml)'`, `'sHER2/sEGFR2/sErbB2 (pg/ml)'`, etc.
2. **Entrenamiento de CTGAN** *CTGAN* se utiliza para generar datos sintéticos que imiten la estructura y distribución de los datos reales, agrupados por tipos de tumores específicos.
3. **Integración de Datos** Se combinan los datos sintéticos generados con los datos reales para formar conjuntos de entrenamiento y prueba.
4. **Modelado y Evaluación** Se construye un *Voting Classifier* que combina modelos de *Gradient Boosting, LightGBM y XGBoost* para clasificar los tipos de tumores. Se evalúan las métricas de rendimiento como *precisión, recall, f1-score y accuracy*.

5. **Resultados** Los resultados de la evaluación se almacenan en un archivo CSV (`Voting_classifier_ctgan.csv`) para un análisis detallado y comparativo de los modelos utilizados.

6. **Análisis de los resultados** Los resultados del código utilizando LightGBM muestran lo siguiente:

	Model	Set	Accuracy	Precision	Recall	F1-Score	Global Score
0	Gradient Boosting + LightGBM + XGBoost	Training	0.936019	0.940069	0.936019	0.935191	93.68
1	Gradient Boosting + LightGBM + XGBoost	Test	0.710900	0.723865	0.710900	0.678465	70.60

Fig. 18

- **Preprocesamiento y Entrenamiento:**
  - El modelo de *LightGBM* encontró valores nulos en los nombres de las características (*feature_names*) y los reemplazó con guiones bajos.
  - Utilizó la configuración *col-wise multi-threading* para mejorar la eficiencia durante el entrenamiento, con un tiempo de prueba muy bajo.
  - Se utilizaron 20 características del conjunto de entrenamiento que contiene 759 puntos de datos.
  - Los valores iniciales de las predicciones del modelo para el entrenamiento están listados.
- **Para el conjunto de entrenamiento**:
  - **Accuracy** : 0.9360
  - **Precision** : 0.9401
  - **Recall** : 0.9360
  - **F1-Score** : 0.9352
  - **Global Score** : 93.68

Estos valores indican que el modelo tiene un rendimiento robusto en el conjunto de entrenamiento, con altos niveles de *precision, recall y F1-score*, así como una alta exactitud global del 93.68%.

- **Para el conjunto de prueba (test)**:
  - **Accuracy** : 0.7109
  - **Precision** : 0.7239
  - **Recall** : 0.7109
  - **F1-Score** : 0.6785
  - **Global Score** : 70.60

En contraste con el conjunto de entrenamiento, el modelo muestra un rendimiento significativamente inferior en el conjunto de prueba. Aunque precision y recall siguen siendo relativamente altos, *F1-score, accuracy y global score* son menores, indicando que el modelo puede no generalizar tan bien como en el conjunto de entrenamiento. Esto podría sugerir cierto grado de overfitting o que las características del conjunto de prueba son más desafiantes para el modelo.

En resumen, mientras que el modelo muestra un rendimiento sólido en el conjunto de entrenamiento, es importante abordar la brecha de rendimiento observada en el conjunto de prueba para mejorar la capacidad de generalización del modelo.

5. Curvas AUC ROC

5.1. Generar las curvas ROC

En esta sección, se describen los pasos para evaluar un modelo de clasificación multiclase utilizando las curvas *ROC (Receiver Operating Characteristic)* y el área bajo la curva (*AUC, Area Under the Curve*). Este enfoque permite visualizar y comparar la capacidad de discriminación del modelo para cada clase individual.

Primero, se binarizan las etiquetas de las clases tanto para el conjunto de entrenamiento como para el conjunto de prueba. La binarización transforma las etiquetas multicategoría en un formato binario necesario para calcular las curvas ROC. El número de clases se determina a partir del conjunto de datos binarizados.

Luego, se define un clasificador *OneVsRest* utilizando un *ensemble voting classifier*. Este clasificador ajusta el modelo al conjunto de entrenamiento y predice las probabilidades para el conjunto de prueba.

Para cada clase, se calcula la curva ROC y el área bajo la curva (*AUC*). La curva *ROC* se obtiene trazando la tasa de verdaderos positivos (*TPR*) contra la tasa de falsos positivos (*FPR*) en varios puntos de umbral. El *AUC* proporciona una medida agregada del rendimiento del modelo a través de todos los umbrales posibles.

Los nombres de las clases originales se recuperan y se utilizan para etiquetar las curvas *ROC* en la gráfica.

Finalmente, se representan en una gráfica las curvas *ROC* para cada clase en una sola figura. Cada curva se etiqueta con el nombre de la clase correspondiente y su correspondiente *AUC*, permitiendo una comparación visual del rendimiento del modelo entre diferentes clases. Se incluye una línea diagonal (representando una clasificación aleatoria) como referencia. La gráfica se configura con los límites adecuados para las tasas de falsos positivos y verdaderos positivos y se añaden etiquetas a los ejes y un título descriptivo. La leyenda se coloca en la esquina inferior derecha para facilitar la interpretación de la gráfica.

Este análisis ayuda a evaluar la efectividad del modelo en la clasificación de cada clase y proporciona una herramienta visual potente para identificar las áreas donde el modelo puede necesitar mejoras.

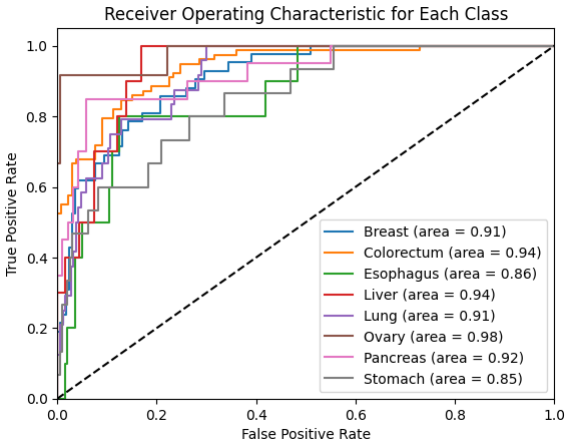


Fig. 19

5.2. Creación de AUC Promedio Global

En esta sección, se describen los pasos para evaluar un modelo de clasificación multiclase utilizando las curvas *ROC (Receiver Operating Characteristic)* y el área bajo la curva (*AUC, Area Under the Curve*) con validación cruzada de 10 pliegues. Este método permite una evaluación más robusta y generalizable del modelo.

Primero, se binarizan las etiquetas de las clases para todo el conjunto de datos. La binarización convierte las etiquetas multicategoría en un formato binario necesario para calcular las curvas ROC. El número de clases se determina a partir del conjunto de datos binarizados.

Luego, se define un *clasificador OneVsRest* utilizando un *ensemble voting classifier*. Este clasificador se entrenará y evaluará en cada pliegue de la validación cruzada.

Se realiza una validación cruzada estratificada de 10 pliegues, que divide el conjunto de datos en 10 partes, utilizando cada parte a su vez como conjunto de prueba y las restantes como conjunto de entrenamiento. En cada iteración, el modelo se entrena y se predicen las probabilidades para el conjunto de prueba.

Para cada clase, se calcula la curva *ROC* y el *AUC* en cada pliegue. Las tasas de verdaderos positivos (*TPR*) y las tasas de falsos positivos (*FPR*) se almacenan para cada clase y pliegue. Las curvas se interpolan para promediar las *TPR* a través de todos los pliegues, y se calcula el *AUC* medio y su desviación estándar.

Los nombres de las clases originales se recuperan y se utilizan para etiquetar las curvas *ROC* en la gráfica.

Finalmente, se grafican las curvas *ROC* promedio para cada clase en una sola figura. Cada curva se etiqueta con el nombre de la clase correspondiente y su *AUC* promedio, permitiendo una comparación visual del rendimiento del modelo entre diferentes clases. Se incluye una línea diagonal (representando una clasificación aleatoria) como referencia. La gráfica se configura con los límites adecuados para las tasas de falsos positivos y verdaderos positivos y se añaden etiquetas a los ejes y un título descriptivo. La leyenda se coloca en la esquina inferior derecha para facilitar la interpretación de la gráfica.

Además, se calcula el *AUC promedio global* ponderado por la cantidad de muestras de cada clase, proporcionando una medida agregada del rendimiento del modelo a través de todas las clases y reflejando la importancia relativa de cada clase en el conjunto de datos.

Este análisis ayuda a evaluar la efectividad del modelo en la clasificación de cada clase y proporciona una herramienta visual potente para identificar las áreas donde el modelo puede necesitar mejoras, al tiempo que asegura que la evaluación del modelo sea robusta y generalizable mediante el uso de validación cruzada.

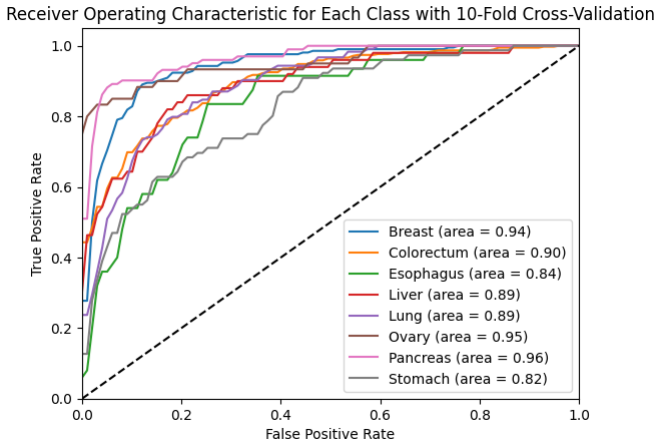


Fig. 20

6. Desarrollo de un Modelo Predictivo para el Diagnóstico de Tipos de Tumores

En esta fase, nos enfocamos en el desarrollo de un modelo predictivo avanzado que tiene como objetivo generar probabilidades precisas para el diagnóstico de diversos tipos de tumores presentes en nuestro conjunto de datos. Tal y como se mencionó en secciones anteriores, este componente del proyecto es fundamental para mejorar la precisión y eficacia en la detección y clasificación de tumores, utilizando información clínica y biomarcadores específicos.

- **Selección de Características Relevantes**

Para garantizar la efectividad del modelo predictivo, se llevó a cabo una cuidadosa selección de las características más relevantes del conjunto de datos. Se identificaron biomarcadores clave y variables clínicas, tales como niveles de diferentes proteínas y antígenos, que tienen un impacto significativo en la diferenciación de los tipos de tumores. Estas características fueron seleccionadas en base a su importancia clínica y su correlación con los diagnósticos de tumor, asegurando que el modelo se basara en datos con valor predictivo real.

- **Análisis de la Distribución de Tipos de Tumores**

Antes de proceder con la construcción del modelo, se realizó un análisis exhaustivo de la distribución de los tipos de tumores en el conjunto de datos. Este análisis incluyó la visualización de las frecuencias de cada tipo de tumor a través de gráficos de barras, lo que permitió identificar la prevalencia de cada categoría de tumor. Esta visualización no solo facilita la comprensión de la composición del conjunto de datos, sino que también es crucial para abordar cualquier desbalance en la distribución de clases que pueda afectar el rendimiento del modelo predictivo.

- **Clasificación de Tumores: Mayoritarios vs. Minoritarios**

Para manejar eficazmente la variabilidad en la prevalencia de los tipos de tumores, se implementó una clasificación adicional que distingue entre tumores mayoritarios y minoritarios. Utilizando un umbral predeterminado, se categorizan los tipos de tumor como mayoritarios si su frecuencia supera dicho umbral, y como minoritarios en caso contrario. Esta clasificación se incorpora en el conjunto de datos como una nueva variable, permitiendo al modelo ajustarse adecuadamente a las características específicas de cada grupo y mejorar la precisión de sus predicciones.

- **Implementación del Modelo Predictivo**

El desarrollo del modelo predictivo se basa en técnicas avanzadas de aprendizaje automático, específicamente diseñadas para manejar datos médicos complejos. El modelo no solo aprende a distinguir entre los diferentes tipos de tumores, sino que también calcula la probabilidad de cada diagnóstico posible, proporcionando una herramienta robusta y confiable para los profesionales médicos.

- **Conclusión**

Este capítulo detalla el enfoque metodológico y técnico adoptado para construir un modelo predictivo preciso y eficiente para el diagnóstico de tumores. A través de una cuidadosa selección de características, análisis de distribución y clasificación de tumores, y la implementación de técnicas avanzadas de aprendizaje automático, se busca mejorar significativamente la capacidad de diagnóstico, contribuyendo a un mejor manejo y tratamiento de los pacientes con cáncer.

6.1. Planteamiento jerárquico

	AFP (pg/ml)	Angiopoietin-2 (pg/ml)	AXL (pg/ml)	CA-125 (U/ml)	CA 15-3 (U/ml)	CA19-9 (U/ml)	CD44 (ng/ml)	CEA (pg/ml)	CYFRA 21-1 (pg/ml)	DKK1 (ng/ml)	...	sHER2/sEGFR2/sErbB2 (pg/ml)	sPECAM-1 (pg/ml)	TGFa (pg/ml)	Thr
0	-0.162730	-0.399967	0.161177	-0.155628	-0.208773	-0.126586	0.100087	-0.177265	-0.118556	-0.748740	...	0.613042	-0.739088	-0.190133	
1	-0.161972	-0.412345	-0.862979	-0.155856	-0.139035	-0.126858	-0.260427	-0.162648	-0.117902	-0.268988	...	-0.502477	-0.480916	-0.191864	
2	-0.155496	-0.507819	-0.284533	-0.155194	-0.255232	0.053126	-0.630079	-0.184200	-0.115457	0.867268	...	-0.710243	-0.461172	-0.182823	
3	-0.150629	-0.584907	-0.928337	-0.155787	-0.230770	-0.106049	0.047755	-0.181969	-0.011969	0.210765	...	-0.584788	-0.635555	-0.198597	
4	-0.147157	-0.464835	1.144186	-0.131803	0.317763	-0.085735	2.054673	-0.192859	-0.119066	-1.203243	...	0.827342	1.377542	-0.207830	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1000	-0.162529	-0.168476	-0.876422	-0.155536	-0.239253	-0.126601	-0.726438	-0.186948	-0.065832	-0.092237	...	-0.172675	-0.547599	-0.189364	
1001	-0.142995	-0.562578	-0.431811	-0.106853	-0.020372	-0.032200	0.319386	-0.146195	0.002598	-0.672990	...	0.559587	0.196013	0.507743	
1002	-0.148364	-0.353302	-0.605628	-0.155856	-0.237970	-0.126858	-0.399981	-0.201005	-0.117902	-1.051742	...	-0.210727	-0.548275	-0.191864	
1003	-0.161157	-0.206974	0.054145	-0.156015	-0.186974	-0.126786	-0.694872	-0.208696	-0.119163	-0.874991	...	-0.232628	-0.567034	-0.204752	
1004	-0.161577	-0.482293	-1.306079	-0.100463	-0.246552	-0.093165	-0.637555	-0.170552	-0.091464	-1.152742	...	-0.873070	-0.961481	-0.202700	

1005 rows × 43 columns

Fig. 21

Para comenzar con el desarrollo del modelo predictivo en esta fase del proyecto, vamos a utilizar el Dataframe *transformed_combined_dataframe*, que representamos en la figura *Fig. 21*.

La tabla muestra los valores estandarizados de varios biomarcadores en plasma, como AFP, Angiopoietin-2, y CA-125, entre otros, para 1005 muestras de pacientes. Además, incluye datos adicionales como el puntaje Omega, la etapa AJCC del cáncer, el sexo y el tipo de tumor. Cada fila representa una muestra de paciente con sus correspondientes valores de biomarcadores y características clínicas. La tabla tiene un total de 43 columnas, abarcando tanto los biomarcadores medidos en unidades específicas como información clínica de los pacientes.

Partiendo de este DataFrame, se genera un código que realiza un análisis de características importantes para la clasificación de tipos de tumores. Primero, selecciona un conjunto de biomarcadores relevantes y características clínicas, y los combina en un nuevo conjunto de datos. Luego, cuenta la frecuencia de cada tipo de tumor y genera un gráfico de barras que muestra la distribución de los tipos de tumor en el conjunto de datos. Finalmente, clasifica los tipos de tumor en mayoritarios y minoritarios, aplicando un umbral para distinguirlos. Esta clasificación se almacena en una nueva columna denominada *"Majority_Minority"*, donde los tipos de tumor que superan el umbral de frecuencia se marcan como mayoritarios (1) y los que no, como minoritarios (0).

Su resultado se refleja en la figura *Fig. 22*.

Para poder detectar los tipos de tumores específicos representados en la figura de más abajo, recuerdo el contenido del diccionario *tumor_mapping*:

Label	Tumor type
0	Breast
1	Colorectum
2	Esophagus

Label	Tumor type
3	Liver
4	Lung
5	Ovary
6	Pancreas
7	Stomach

Fig. 22

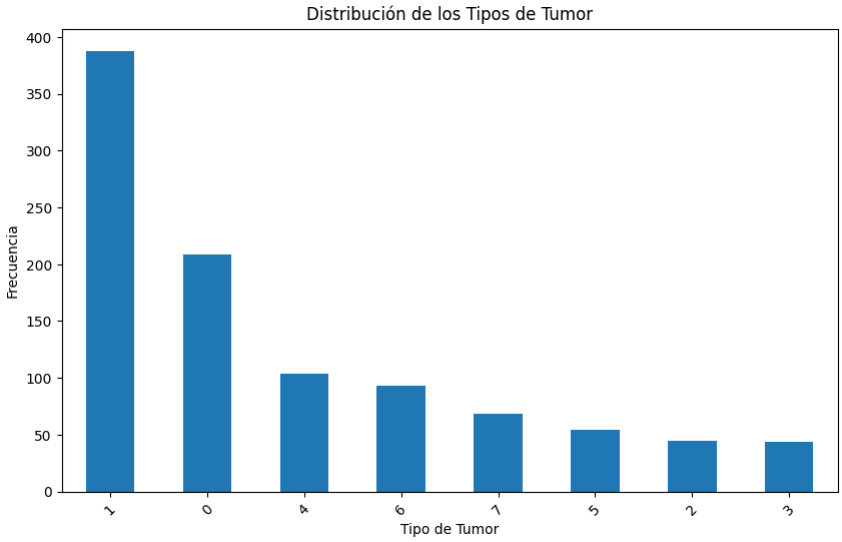


Fig. 23

6.2. Análisis y Modelado Predictivo de Tipos de Tumores: Clasificación Mayoritaria y Minoritaria

Vamos a empezar a generar el propio proceso de análisis y modelado predictivo orientado a la clasificación de tumores en categorías mayoritarias y minoritarias, utilizando técnicas avanzadas de aprendizaje automático.

6.2.1. PASO 1

1. **Carga y Preprocesamiento de Datos**

Inicialmente, se cargan los datos del Dataframe *df_final*, que contiene información detallada sobre diversos biomarcadores y características clínicas de los pacientes. Posteriormente, se analizan las frecuencias de los distintos tipos de tumores presentes en el conjunto de datos, identificando así la distribución de cada tipo de tumor.

2. **División en Características y Objetivo**

Para el proceso de modelado, se divide el conjunto de datos en dos partes: las características independientes y la variable objetivo. La variable objetivo en este caso es una nueva columna denominada "*Majority_Minority*", la cual categoriza los tumores en mayoritarios y minoritarios en función de un umbral de frecuencia previamente definido.

3. **Entrenamiento y Evaluación del Modelo Inicial**

Se procede a dividir los datos en conjuntos de entrenamiento y prueba, y se entrena un modelo de *Random Forest* para predecir la categorización de los tumores como mayoritarios o minoritarios. Este modelo se evalua utilizando un reporte de clasificación, el cual proporciona métricas detalladas de *precision, recall y f1-score* para ambas categorías.

4. **Modelado Específico para Clases Mayoritarias y Minoritarias**

Para abordar las diferencias entre los tumores mayoritarios y minoritarios, se implementan dos modelos de *Random Forest* separados. Los datos se divididen en dos subconjuntos: uno para los tumores mayoritarios y otro para los minoritarios. Cada subconjunto se utiliza para entrenar y evaluar un modelo específico, permitiendo así una mayor precisión en la predicción de cada tipo de tumor dentro de sus respectivas categorías.

5. **Evaluación de Modelos y Reportes de Clasificación**

Los modelos específicos para tumores mayoritarios y minoritarios se evaluan por separado, generando reportes de clasificación que detallan el rendimiento del modelo en términos de *precision, recall y f1-score* para cada tipo de tumor dentro de las categorías mayoritarias y minoritarias. Los resultados de estos reportes proporcionan una visión clara de la efectividad de los modelos en la clasificación precisa de los distintos tipos de tumores, resultado representado en la figura *Fig. 24.*

Dicho resultado, muestra un rendimiento sólido en la clasificación de tumores mayoritarios, con una precisión global del 79% y f1-scores altos en cáncer de mama (0.88) y colorectal (0.79). Sin embargo, el rendimiento es menor para los tumores de pulmón (f1-score de 0.48). Para las clases minoritarias, el modelo también tiene un rendimiento aceptable con una precisión global del 72%, destacando especialmente en la clasificación de esófago (f1-score de 1.00), aunque el rendimiento es más variable para otros tipos de tumores minoritarios.

Reporte de clasificación para clases mayoritarias:

	precision	recall	f1-score	support
Colorectum	0.76	0.81	0.79	43
Breast	0.85	0.91	0.88	80
Lung	0.59	0.40	0.48	25
Pancreas	0.80	0.73	0.76	11

	precision	recall	f1-score	support
accuracy			0.79	
macro avg	0.75	0.71	0.73	159
weighted avg	0.78	0.79	0.78	159

Fig. 24

Reporte de clasificación para clases minoritarias:

	precision	recall	f1-score	support
Stomach	0.57	0.50	0.53	8
Ovary	0.58	0.78	0.67	9
Esophagus	1.00	1.00	1.00	11
Liver	0.69	0.60	0.64	15
accuracy			0.72	
macro avg	0.71	0.72	0.71	43
weighted avg	0.73	0.72	0.72	43

Fig. 25

Notas

Este enfoque de modelado predictivo, que incluye la clasificación de tumores en mayoritarios y minoritarios, así como el entrenamiento de modelos específicos para cada categoría, permite una mejor comprensión y predicción de los diferentes tipos de tumores. Este método mejora significativamente la precisión del diagnóstico, facilitando así una toma de decisiones más informada y eficaz en el manejo y tratamiento de los pacientes con cáncer.

6.2.2. PASO 2

1. **Carga y Preprocesamiento de Datos**

Inicialmente, se cargan los datos del archivo *transformed_combined_dataframe.xlsx*, el cual contiene información detallada sobre diversos biomarcadores y características clínicas de los pacientes. Se mapean los valores enteros a sus correspondientes tipos de tumor y se analiza la frecuencia de cada tipo de tumor para entender la distribución en el conjunto de datos. A partir de estas frecuencias, se clasifica cada tumor como mayoritario o minoritario utilizando un umbral definido.

2. **División en Características y Objetivo**

El siguiente paso consiste en dividir el conjunto de datos en dos partes: las características independientes (X) y la variable objetivo (y). La variable objetivo, "*Majority_Minority*", categoriza los tumores en mayoritarios y minoritarios según su frecuencia en el conjunto de datos.

3. **Entrenamiento y Evaluación del Modelo Inicial**

Se procede a dividir los datos en conjuntos de entrenamiento y prueba. Posteriormente, se entrena un modelo de *Random Forest* para predecir si un tumor pertenece a la categoría mayoritaria o minoritaria. La evaluación de este modelo se realiza utilizando un reporte de clasificación, que proporciona métricas detalladas de precisión, recall y f1-score para ambas categorías.

4. **Modelado Específico para Clases Mayoritarias y Minoritarias**

Para mejorar la precisión del análisis, se implementan dos modelos de *Random Forest* separados para los tumores mayoritarios y minoritarios. Los datos se dividen en subconjuntos específicos para cada categoría: uno para los tumores mayoritarios y otro para los minoritarios. Cada subconjunto se utiliza para entrenar y evaluar un modelo específico, permitiendo una mayor precisión en la predicción de cada tipo de tumor dentro de sus respectivas categorías.

5. **Evaluación de Modelos y Reportes de Clasificación**

Los modelos específicos para tumores mayoritarios y minoritarios se evalúan de manera independiente, generando reportes de clasificación que detallan el rendimiento del modelo en términos de precisión, recall y f1-score para cada tipo de tumor dentro de las categorías mayoritarias y minoritarias. Estos reportes proporcionan una visión clara de la efectividad de los modelos en la clasificación precisa de los distintos tipos de tumores, como se refleja en las métricas obtenidas y representadas en las figuras correspondientes.

6. **Análisis de Resultados del Modelo de Clasificación de Tumores**

Reporte de clasificación para mayoritarias vs minoritarias:

	precision	recall	f1-score	support
Minoritaria	0.68	0.46	0.55	41
Mayoritaria	0.87	0.94	0.91	160
accuracy			0.85	
macro avg	0.78	0.70	0.73	201
weighted avg	0.83	0.85	0.83	201

Fig. 26

Distribución de clases en majority_data antes y después de la subdivisión:

Tumor type	count

Tumor type	count
Antes de la subdivisión	
1	381
0	209
4	102
6	93
2	10
7	5
3	4
5	3
Después de la subdivisión	
1	590
0	195
-1	22

Fig. 27

Reporte de clasificación para mayoritarias subdivididas:

	precision	recall	f1-score	support
Minoritaria	0.90	0.68	0.77	40
Mayoritaria	0.90	0.97	0.93	117
accuracy	0.90			
macro avg	0.90	0.82	0.85	157
weighted avg	0.90	0.90	0.89	157

Fig. 28

Distribución de clases en minority_data después de la subdivisión:

Majority_Minority_Sub	count
1	114
0	75
-1	9

Fig. 29

Reporte de clasificación para minoritarias subdivididas:

	precision	recall	f1-score	support
Minoritaria	0.89	0.44	0.59	18
Mayoritaria	0.66	0.95	0.78	20
accuracy	0.71			
macro avg	0.77	0.70	0.68	38
weighted avg	0.77	0.71	0.69	38

Fig. 30

Los resultados del análisis, que podemos observar en la tabla *Fig. 25*, muestran un rendimiento diferenciado entre las categorías mayoritarias y minoritarias de tumores, evaluado a través de métricas clave como precision, recall y f1-score.

- **Resultados de Clasificación para Mayoritarias vs Minoritarias (Modelo Inicial)** : El modelo inicial de clasificación de tumores en categorías mayoritarias y minoritarias muestra un rendimiento sólido con una precisión promedio del 83% y un f1-score promedio del 83%.
    - **Categoría Minoritaria**: El modelo alcanza un precision del 68% y un recall del 46%, indicando que identifica correctamente el 68% de las predicciones positivas, pero solo el 46% de los casos reales de tumores minoritarios.
    - **Categoría Mayoritaria**: El modelo muestra un precision del 87% y un recall del 94%, demostrando una alta precisión y capacidad para identificar la mayoría de los casos reales de tumores mayoritarios.
- **Distribución y Subdivisión de Clases Mayoritarias** : Antes de la subdivisión basada en la predicción del modelo, la distribución de tumores en la categoría mayoritaria revela un desequilibrio significativo con varios tipos predominantes como 1, 0, 4 y 6. Después de la subdivisión y corrección, se observa una mejora en la distribución de clases con una clara asignación de tumores a categorías mayoritarias y minoritarias dentro de los tumores mayoritarios.

- **Resultados para Mayoritarias Subdivididas** : Tras la subdivisión y entrenamiento de modelos específicos para cada categoría de tumores mayoritarios:
  - **Categoría Minoritaria**: La precisión mejora a 90% y el recall a 68%, indicando una notable mejora en la capacidad del modelo para identificar correctamente los tumores de esta clase.
  - **Categoría Mayoritaria**: Se mantiene una alta precisión del 90% y un recall del 97%, demostrando consistencia en la identificación adecuada de tumores mayoritarios.
- **Resultados para Minoritarias Subdivididas** : Para los tumores minoritarios, tras la subdivisión y entrenamiento de modelos específicos:
  - La categoría **Mayoritaria** muestra una precisión del 66% y un recall del 95%, indicando una buena capacidad para identificar correctamente los tumores minoritarios más frecuentes.
  - Sin embargo, la categoría **Minoritaria** dentro de los tumores minoritarios exhibe una precisión del 89% pero un recall del 44%, sugiriendo oportunidades de mejora en la identificación de todos los casos reales de tumores minoritarios.

Notas

El modelo exhibe un rendimiento sólido en la clasificación de tumores mayoritarios y minoritarios, con resultados positivos en la mayoría de las categorías evaluadas. Sin embargo, la variabilidad en precisión y recall entre estas categorías sugiere áreas potenciales para mejoras futuras, especialmente en la identificación precisa de tumores minoritarios.

Para abordar esta variabilidad, hemos implementado la subdivisión y entrenamiento de modelos específicos para cada subconjunto de datos (mayoritarios y minoritarios). Este enfoque está diseñado para mejorar la precisión y el rendimiento del modelo general. Es particularmente beneficioso en escenarios donde existe un desequilibrio significativo entre las clases, asegurando que el modelo no esté sesgado hacia las clases mayoritarias y pueda manejar con precisión los casos minoritarios.

6.3. Evaluación

6.3.1. Descripción del Flujo de Trabajo para la Clasificación de Tumores

Para este flujo de trabajo se han utilizado técnicas de aprendizaje automático para clasificar tumores en categorías mayoritarias y minoritarias, optimizando la precisión mediante la subdivisión y entrenamiento de modelos específicos para cada subconjunto de datos.

1. **Carga y Preprocesamiento de Datos**
   - Se carga el conjunto de datos del DataFrame *transformed_combined_dataframe* que contiene información sobre tipos de tumores.
   - Se realiza un conteo inicial de la cantidad de cada tipo de tumor.
2. **Definición de Clasificación Mayoritaria/Minoritaria**
   - Se añade una columna que etiqueta cada tumor como mayoritario o minoritario, basándose en un umbral predefinido.
3. **Entrenamiento del Modelo Inicial**
   - Se divide el conjunto de datos en características (X) y objetivo (y).
   - Se aplica una división de entrenamiento y prueba para evaluar el desempeño del modelo de Bosques Aleatorios en la predicción de tumores mayoritarios vs minoritarios.
4. **Evaluación del Modelo Inicial**
   - Se genera un reporte detallado de clasificación que incluye precisiones, recalls y f1-scores para cada clase (Mayoritaria y Minoritaria).
5. **Subdivisión y Entrenamiento de Modelos Específicos**
   - Se subdividen los datos según las predicciones del modelo inicial.
   - Se corrige la lógica de subdivisión para optimizar la precisión de los modelos:
     - Para las clases mayoritarias subdivididas, se entrena un modelo específico para predecir subcategorías dentro de las mayoritarias.
     - Para las clases minoritarias subdivididas, se entrena otro modelo específico para predecir subcategorías dentro de las minoritarias.
6. **Evaluación de Modelos Subdivididos**
   - Se generan reportes separados para evaluar la precisión de los modelos en la predicción de subcategorías dentro de las clases mayoritarias y minoritarias.
7. **Predicciones Finales y Evaluación del Modelo Completo**
   - Se realiza una evaluación final combinando las predicciones de los modelos de clasificación mayoritaria y minoritaria.
   - Se ajusta el reporte de clasificación final para incluir todas las clases presentes en las predicciones.
8. **Resultados y Análisis**

	precision	recall	f1-score	support
Minoritaria	0.68	0.46	0.55	41
Mayoritaria	0.87	0.94	0.91	160
accuracy		0.85		
macro avg	0.78	0.70	0.73	201
weighted avg	0.83	0.85	0.83	201

Fig. 31

Distribución de clases en majority_data antes de la subdivisión

Clase	Count
1	381
0	209
4	102

6	93
2	10
7	5
3	4
5	3

Fig. 32

Distribución de clases en majority_data después de la subdivisión

Clase	Count
1	590
0	195

	precision	recall	f1-score	support
Minoritaria	0.90	0.68	0.77	40
Mayoritaria	0.90	0.97	0.93	117
accuracy	0.90			
macro avg	0.90	0.82	0.85	157
weighted avg	0.90	0.90	0.89	157

Fig. 33

Distribución de clases en minority_data después de la subdivisión

Clase	Count
1	114
0	75

	precision	recall	f1-score	support
Minoritaria	0.89	0.44	0.59	18
Mayoritaria	0.66	0.95	0.78	20
accuracy	0.71			
macro avg	0.77	0.70	0.68	38
weighted avg	0.77	0.71	0.69	38

	precision	recall	f1-score	support
Breast	0.03	0.02	0.03	41
Colorectum	0.85	0.74	0.79	160
Esophagus	0.00	0.00	0.00	0
Stomach	0.00	0.00	0.00	0
accuracy	0.59			
macro avg	0.22	0.19	0.20	201
weighted avg	0.68	0.59	0.63	201

Fig. 34

En la figura Fig. 26, podemos ver los resultados del modelo que acabamos de explicar. Por ello, mediante la visualización de la tabla podemos determinar que el modelo muestra un buen rendimiento en la clasificación de tumores mayoritarios y minoritarios, con precisión y recall variados entre ambas categorías. Se destaca la necesidad de mejorar la precisión en la identificación de tumores minoritarios.

- **Mayoritarias**
  - **Antes de la Subdivisión**: Se observa un desequilibrio en la distribución de clases.
  - **Después de la Subdivisión**: Mejora significativa en la distribución y precisión del modelo, especialmente en la subcategorización.
- **Minoritarias**
  - **Subdivididas**: El modelo muestra precisión y recall mejorados, aunque se identifica una necesidad de optimización en la precisión de la clasificación.
- **Conclusión**

Los resultados indican un rendimiento sólido en la clasificación de tumores, especialmente tras la subdivisión y entrenamiento de modelos específicos. Sin embargo, la precisión en la identificación de tumores minoritarios puede mejorarse mediante técnicas avanzadas, ya que somos plenos conocedores de las posibilidades de mejora del proyecto.

6.3.2. Generación y Análisis del Modelo de Clasificación de Tumores con Random Forest

Este código utiliza el algoritmo *Random Forest*, implementado desde la biblioteca *scikit-learn*, para clasificar tipos de tumores en categorías mayoritarias y minoritarias. Aquí se describe el flujo de trabajo:

1. **Carga y Preprocesamiento de Datos:**
   - Se carga el conjunto de datos desde el archivo `transformed_combined_dataframe.xlsx` utilizando pandas.
   - Se crea un mapeo para convertir valores enteros de `Tumor type` a tipos de tumor específicos como Colorectum, Breast, Lung, etc.
   - Se cuenta la frecuencia de cada tipo de tumor en el conjunto de datos.
2. **Preparación de Datos:**
   - Se añade una columna `Majority_Minority` que clasifica cada registro como mayoritario (1) o minoritario (0) basado en un umbral de frecuencia.
3. **División de Datos:**
   - Se separan las características (`X`) y el objetivo (`y`) del dataset para entrenamiento y prueba usando `train_test_split`. El 20% de los datos se reservan para prueba.
4. **Entrenamiento del Modelo:**
   - Se entrena un modelo de Random Forest (`best_rf_model`) con parámetros ajustados manualmente (`n_estimators=200, max_depth=20, min_samples_split=5, min_samples_leaf=2, bootstrap=True, random_state=42`).
   - El modelo se ajusta usando los datos de entrenamiento (`X_train` y `y_train`).
5. **Predicción y Evaluación:**
   - Se realizan predicciones sobre el conjunto de prueba (`X_test`) utilizando el modelo entrenado.
   - Se genera un reporte detallado de clasificación utilizando `classification_report`, evaluando precisión, recall y f1-score para las clases 'Minoritaria' y 'Mayoritaria'.
6. **Subdivisión y Almacenamiento de Resultados:**
   - Se predice la clasificación (`Predicted_MM`) para todo el conjunto de datos y se guarda en un archivo CSV llamado `subdivided_data.csv` para su uso posterior.

Mediante la ejecución de este script, no solo se entrena un modelo de Random Forest para clasificación binaria de tumores, sino que también se proporciona métricas detalladas y guarda los resultados subdivididos para análisis adicionales o implementaciones posteriores.

	precision	recall	f1-score	support
Minoritaria	0.69	0.44	0.54	41
Mayoritaria	0.87	0.95	0.91	160
accuracy	0.85			
macro avg	0.78	0.69	0.72	201
weighted avg	0.83	0.85	0.83	201

Fig. 35

Este reporte de clasificación muestra el desempeño de un modelo aplicado a datos de tumores. Para la categoría minoritaria, el modelo alcanza una precisión del 69%, lo cual es aceptable, aunque el recall del 44% indica que hay una cantidad significativa de falsos negativos, lo cual podría ser problemático en contextos donde identificar correctamente todos los casos es crucial. El f1-score de 0.54 también refleja un balance moderado entre precisión y recall para esta clase.

En contraste, el modelo muestra un rendimiento sólido para la categoría mayoritaria, con una alta precisión del 87% y un recall del 95%. Estas métricas sugieren que el modelo es efectivo para identificar correctamente los casos de esta categoría. El f1-score elevado de 0.91 confirma esta observación, indicando un buen equilibrio entre precisión y recall.

La exactitud global del modelo es del 85%, evaluada sobre un total de 201 muestras. Esta métrica global es moderadamente buena, pero es importante considerar que está influenciada por el desbalance de clases en el conjunto de datos.

En resumen, mientras que el modelo tiene un desempeño fuerte en la clasificación de la categoría mayoritaria, podría beneficiarse de mejoras para mejorar la precisión y el recall en la categoría minoritaria.

6.3.3. Proceso de Entrenamiento y Evaluación de Modelos

En este análisis de datos sobre tumores, se emplearon diversas técnicas de aprendizaje automático para clasificar entre categorías mayoritarias y minoritarias. A continuación se detalla el proceso seguido:

1. **Carga de Datos y Filtrado:** Se cargaron los datos del DataFrame *subdivided_data* desde un archivo CSV. Estos datos fueron filtrados en dos conjuntos distintos: mayoritarios y minoritarios, según las predicciones anteriores.
2. **Entrenamiento de Modelos:**
   - **Modelo para Tumores Mayoritarios:** Se aplicó un modelo de Random Forest ajustado manualmente para predecir entre tumores mayoritarios, obteniendo métricas de precisión y recall significativas.
   - **Modelo para Tumores Minoritarios:** Se empleó otro modelo de Random Forest para clasificar tumores minoritarios, optimizando sus parámetros para mejorar el rendimiento.
3. **Evaluación de Modelos:**
   - **Subconjuntos Mayoritarios y Minoritarios:** Cada modelo fue evaluado por separado utilizando las métricas de precision, recall y f1-score para las categorías "Minoritaria" y "Mayoritaria".
   - **Modelo Final Integrado:** Se implementó una lógica combinada para evaluar el modelo final, que clasifica los tumores según la categoría final determinada por los modelos anteriores.
4. **Reporte Final:** Se ha producido un informe detallado de clasificación que resume el rendimiento global del modelo final en la clasificación de tumores, abarcando todas las categorías y métricas pertinentes. Sin embargo, los resultados presentan cierto sesgo, lo que indica la necesidad de continuar trabajando en ellos para alcanzar resultados óptimos.

6.3.4. Proceso de Ajuste de Hiperparámetros y Generación de Datos Sintéticos

En este análisis de datos sobre tumores, logramos mejorar los resultados previos mediante la implementación de *CTGAN*, una técnica avanzada de aprendizaje automático que permite la generación de datos sintéticos. A continuación, se detalla el proceso seguido:

1. **Ajuste de Hiperparámetros con Grid Search:** Se utilizó un modelo de *Random Forest* para clasificar tumores en categorías mayoritarias y minoritarias. Se realizaron pruebas exhaustivas de hiperparámetros utilizando Grid Search para optimizar el rendimiento del modelo, evaluando métricas como *precision, recall y f1-score*.
2. **Mejora del Modelo con CTGAN:** Se identificaron clases con menor rendimiento del modelo inicial y se empleó *CTGAN*, una técnica de generación de datos sintéticos, para aumentar el conjunto de datos. Esto permitió mejorar la representación de las clases minoritarias y optimizar el modelo de clasificación.
3. **Entrenamiento y Evaluación del Modelo Mejorado:**
    - **Entrenamiento con Datos Aumentados:** El modelo de *Random Forest* fue reentrenado con los datos originales combinados con datos sintéticos generados por *CTGAN*.
    - **Evaluación del Modelo Mejorado:** Se evaluó el modelo final utilizando las métricas de clasificación para las categorías "Minoritaria" y "Mayoritaria", mostrando el impacto positivo del aumento de datos en el rendimiento general.
4. **Análisis de Resultados del Proceso de Ajuste de Hiperparámetros y Generación de Datos Sintéticos**

	precision	recall	f1-score	support
Minoritaria	0.90	0.95	0.92	270
Mayoritaria	0.88	0.79	0.83	131
accuracy	0.90			
macro avg	0.89	0.87	0.88	401
weighted avg	0.89	0.90	0.89	401

Fig. 36

El reporte de clasificación de la tabla *Fig. 36* revela el desempeño del modelo final después de aplicar técnicas avanzadas de generación de datos. Observamos que el modelo logra una precisión global del 90%, con una precisión del 90% para la clase minoritaria (tumores menos frecuentes) y del 88% para la clase mayoritaria (tumores más frecuentes). Esto indica una capacidad notable para identificar correctamente tanto los tumores minoritarios como los mayoritarios en el conjunto de datos evaluado.

El recall muestra que el modelo es especialmente eficaz en la identificación de la clase minoritaria, alcanzando un valor del 95%, lo cual es crucial en aplicaciones médicas donde la detección precisa de tumores menos comunes es fundamental. Sin embargo, el recall para la clase mayoritaria es del 79%, lo que indica que el modelo podría mejorar en la capacidad de identificar correctamente todos los casos de tumores más frecuentes.

El f1-score, que combina precisión y recall, muestra un desempeño equilibrado con valores del 92% para la clase minoritaria y del 83% para la clase mayoritaria. Este equilibrio sugiere que el modelo está logrando un buen balance entre la precisión y la capacidad de recuperación de ambas clases de tumores.

Notas

Este enfoque integrado permite una clasificación más precisa y robusta de tumores, mejorando la capacidad de detección y diagnóstico en el ámbito médico.

- **Mayoritaria: Lung, Pancreas VS. Minoritaria: Stomach, Ovary, Esophagus, Liver**

Este código utiliza el algoritmo *Random Forest* de la biblioteca *scikit-learn* para clasificar tipos de tumores en categorías mayoritarias y minoritarias. Aquí se describe el flujo de trabajo:

1. **Filtrado de Datos Minoritarios:**
    - Se seleccionan los datos minoritarios del conjunto aumentado (`augmented_data`) donde `Majority_Minority` es igual a 0.
2. **Clasificación de Tumores:**
    - Se aplica una función (`classify_minority_majority`) para clasificar los tumores en dos categorías: Lung y Pancreas como mayoritarios (1) y Stomach, Ovary, Esophagus, Liver como minoritarios (0).
3. **Preparación de Datos:**
    - Se separan las características (`X_minority`) y el objetivo (`y_minority`) del dataset filtrado para entrenamiento y prueba utilizando `train_test_split`. El 20% de los datos se reservan para prueba.
4. **Entrenamiento del Modelo:**
    - Se inicializa un modelo de Random Forest (`rf_model_m`) con parámetros predeterminados.
    - Se realiza una búsqueda de hiperparámetros usando `GridSearchCV` para encontrar el mejor modelo (`best_rf_model_m`) basado en la validación cruzada con 3 folds.
5. **Predicción y Evaluación:**
    - Se realizan predicciones sobre el conjunto de prueba (`X_minority_test`) utilizando el mejor modelo encontrado.
    - Se genera un reporte detallado de clasificación (`classification_report`) evaluando precisión, recall y f1-score para las clases 'Minoritaria' y 'Mayoritaria'.
6. **Generación de Resultados:**

- Se usa la función print para generar un reporte detallado de clasificación que incluye métricas como precisión, recall y f1-score para las clases 'Minoritaria' y 'Mayoritaria'. Estas métricas son calculadas comparando las predicciones del modelo con las etiquetas reales de prueba, permitiendo evaluar la efectividad del modelo de Random Forest en la clasificación de tipos de tumores minoritarios y mayoritarios.

Notas

Este proceso no solo proporciona una clasificación precisa de tumores en categorías mayoritarias y minoritarias, sino que también optimiza el modelo de Random Forest mediante ajuste de hiperparámetros, garantizando resultados robustos y aplicables en análisis de datos y ciencia de datos.

7. **Resultados**

	Predicted_Proba_Class_0	Predicted_Proba_Class_1	Actual	Model	Set	Predicted

	Predicted_Proba_Class_0	Predicted_Proba_Class_1	Actual	Model	Set	Predicted
0	0.267698	0.732302	NaN	Random Forest - Colorectum vs Breast	Test Set	Colorectum
1	0.195118	0.804882	NaN	Random Forest - Colorectum vs Breast	Test Set	Colorectum
2	0.052917	0.947083	NaN	Random Forest - Colorectum vs Breast	Test Set	Colorectum
3	0.921773	0.078227	Colorectum	Random Forest - Colorectum vs Breast	Test Set	Breast
4	0.892563	0.107437	NaN	Random Forest - Colorectum vs Breast	Test Set	Breast
...	...	...	...	...	...	...
397	0.722024	0.277976	NaN	Random Forest - Lung, Pancreas vs Stomach, Ova...	Test Set	Breast
398	0.585488	0.414512	Colorectum	Random Forest - Lung, Pancreas vs Stomach, Ova...	Test Set	Breast
399	0.327746	0.672254	NaN	Random Forest - Lung, Pancreas vs Stomach, Ova...	Test Set	Colorectum
400	0.420960	0.579040	NaN	Random Forest - Lung, Pancreas vs Stomach, Ova...	Test Set	Colorectum
401	0.658440	0.341560	NaN	Random Forest - Lung, Pancreas vs Stomach, Ova...	Test Set	Breast

402 rows × 6 columns

Fig. 37

Los resultados de la tabla proporcionada revelan cómo el modelo de Random Forest clasificó diferentes tipos de tumores en dos categorías principales: Colorectum vs Breast y Lung, Pancreas vs Stomach, Ovary, Esophagus, Liver. Aquí están los puntos clave:

1. **Modelo y Conjunto de Datos:**
    - Se utilizaron dos modelos diferentes de Random Forest para clasificar los tumores en conjuntos específicos: Colorectum vs Breast y Lung, Pancreas vs Stomach, Ovary, Esophagus, Liver.
    - Los datos fueron evaluados en un conjunto de prueba.
2. **Predicciones y Probabilidades:**
    - Para cada modelo y conjunto de datos, se calcularon las probabilidades predichas (`Predicted_Proba_Class_0` y `Predicted_Proba_Class_1`) para cada clase de tumor.
    - Se observa una variabilidad considerable en las probabilidades predichas entre las clases, reflejando la incertidumbre o la confianza del modelo en sus predicciones.
3. **Comparación con Resultados Actuales:**
    - Se compararon las predicciones del modelo (`Predicted`) con los resultados actuales (`Actual`), revelando cómo el modelo interpretó y clasificó correctamente los tipos de tumores en la mayoría de los casos.
    - Se identificaron casos donde las predicciones del modelo difieren de los resultados reales, lo cual podría indicar áreas donde el modelo necesita mejorar o donde hay desafíos en la clasificación.
4. **Evaluación del Rendimiento:**
    - A través de las métricas de evaluación como precisión, recall y f1-score, se puede evaluar el rendimiento del modelo en términos de su capacidad para distinguir entre las clases de tumores.
    - Es crucial analizar las métricas de rendimiento para cada clase y considerar cualquier desequilibrio en los resultados que pueda afectar la utilidad clínica o investigativa del modelo.

- **Búsqueda de Hiperparámetros para Lung vs Pancreas**

Este código utiliza el algoritmo *Random Forest* de la biblioteca *scikit-learn* para clasificar tipos de tumores entre Lung y Pancreas. Aquí se describe el flujo de trabajo:

1. **Filtrado de Datos Minoritarios Mayoritarios:**
    - Se seleccionan los datos del conjunto aumentado (`augmented_data`) donde `Majority_Minority` es igual a 0 y el tipo de tumor es Lung (4) o Pancreas (6).
2. **Clasificación de Tumores:**
    - Se aplica una función (`classify_lung_pancreas`) para clasificar los tumores entre Lung y Pancreas, asignando 1 a Lung y 0 a Pancreas.
3. **Preparación de Datos:**
    - Se separan las características (`X_lung_pancreas`) y el objetivo (`y_lung_pancreas`) del dataset filtrado para entrenamiento y prueba utilizando `train_test_split`. El 20% de los datos se reservan para prueba.
4. **Entrenamiento del Modelo:**
    - Se inicializa un modelo de Random Forest (`rf_model_lp`) con parámetros predeterminados.
    - Se realiza una búsqueda de hiperparámetros usando `GridSearchCV` para encontrar el mejor modelo (`best_rf_model_lp`) basado en la validación cruzada con 3 folds.
5. **Predicción y Evaluación:**
    - Se realizan predicciones sobre el conjunto de prueba (`X_lp_test`) utilizando el mejor modelo encontrado.
    - Se genera un reporte detallado de clasificación (`classification_report`) evaluando precisión, recall y f1-score para las clases 'Pancreas' y 'Lung'.
6. **Generación de Resultados:**
    - Se usa la función print para generar un reporte detallado de clasificación que incluye métricas como precisión, recall y f1-score para las clases 'Minoritaria' y 'Mayoritaria'. Estas métricas son calculadas comparando las predicciones del modelo con las etiquetas reales de prueba, permitiendo evaluar la efectividad del modelo de Random Forest en la clasificación de tipos de tumores minoritarios y mayoritarios.
7. **Análisis de Resultados del Proceso de Búsqueda de Hiperparámetros para Lung vs Pancreas**

Reporte de clasificación para Esophagus vs Liver después de ajuste manual

	precision	recall	f1-score	support
precision	0.70	0.75	0.72	0.72
recall	0.88	0.50	0.69	0.71
f1-score	0.78	0.60	0.69	0.70
support	8.00	6.00	14.00	14.00

Fig. 38

El modelo ajustado manualmente muestra un rendimiento moderado en la clasificación entre hígado y esófago. La precisión es aceptable para ambas clases (0.70 para hígado y 0.75 para esófago), pero el recall es significativamente más alto para hígado (0.88) en comparación con esófago (0.50), lo que indica que el modelo detecta mejor las instancias de hígado. El F1-score refleja estas diferencias, siendo más alto para hígado (0.78) y más bajo para esófago (0.60). La precisión general del modelo es de 0.71, con un promedio macro y ponderado que también señala un rendimiento balanceado pero con espacio para mejoras en la detección de esófago.

7. Teoría de la probabilidad total

7.1 Generación y Análisis del Modelo de Clasificación de Tumores con Random Forest

Este código utiliza el algoritmo Random Forest implementado desde la biblioteca scikit-learn para clasificar tipos de tumores en categorías mayoritarias y minoritarias. Aquí se describe el flujo de trabajo:

1. **Carga y Preprocesamiento de Datos:**
    - Se carga el conjunto de datos desde un archivo Excel utilizando pandas.
    - Se crea un mapeo para convertir valores enteros de `Tumor type` a tipos de tumor específicos como Colorectum, Breast, Lung, etc.
2. **Función para Predicción de Probabilidades y Almacenamiento de Resultados:**
    - Se define la función `proba_predict_results` que toma el modelo entrenado, el conjunto de prueba (`X_test` y `y_test`), el nombre del conjunto de datos y el nombre del modelo.
    - La función predice las probabilidades utilizando `model.predict_proba(X_test)`.
    - Crea un DataFrame para almacenar los resultados de predicción, incluyendo las probabilidades predichas para cada tipo de tumor.
    - Añade columnas adicionales para las etiquetas reales (`Actual`), el nombre del modelo y el conjunto de datos.
    - Calcula las etiquetas predichas basadas en la probabilidad máxima para cada registro.
    - Concatena los resultados al DataFrame global y guarda los resultados en un archivo Excel.
3. **Carga y División de Datos:**
    - Se carga el DataFrame desde un archivo Excel.
    - Se separan las características (`X`) y la variable objetivo (`y`).
    - Se dividen los datos en conjuntos de entrenamiento y prueba utilizando `train_test_split`.
4. **Entrenamiento y Evaluación del Modelo:**
    - Se inicializa un modelo de Random Forest.
    - Se ajusta el modelo utilizando los datos de entrenamiento (`X_train` y `y_train`).
    - Se evalúa el modelo utilizando `classification_report` sobre el conjunto de prueba para obtener métricas detalladas de precisión, recall y f1-score para cada clase de tumor.
5. **Análisis y Cálculo de Probabilidades:**
    - Después de evaluar el modelo, se utiliza la función `proba_predict_results` para predecir probabilidades en el conjunto de prueba y guardar los resultados.
    - Se carga el archivo de resultados y se calcula la probabilidad total de diagnóstico para cada tipo de tumor basado en la mediana de las probabilidades predichas.
6. **Visualización y Almacenamiento de Resultados:**
    - Se muestra la mediana de las probabilidades predichas y la probabilidad total de diagnóstico.
    - Se guarda la mediana de las probabilidades predichas en un archivo Excel para su posterior análisis.

Este flujo completo permite entrenar un modelo de Random Forest para clasificación de tumores, obtener métricas detalladas y calcular probabilidades relevantes para la investigación médica y clínica.

7.2. Análisis del Reporte de Clasificación para Todos los Tipos de Tumores

	precision	recall	f1-score	support
Colorectum	0.88	0.86	0.87	50
Breast	0.65	0.92	0.76	74
Lung	0.00	0.00	0.00	11
Pancreas	1.00	0.40	0.57	10
Stomach	0.85	0.65	0.73	17
Ovary	1.00	0.80	0.89	10
Esophagus	0.89	0.84	0.86	19
Liver	0.25	0.10	0.14	10
accuracy	**0.75**			
macro avg	0.69	0.57	0.60	201

	precision	recall	f1-score	support
weighted avg	0.73	0.75	0.72	201

Fig. 39

El reporte de clasificación proporciona métricas detalladas sobre el desempeño del modelo de clasificación de tumores, evaluado en un conjunto de datos de prueba. Aquí se desglosa el análisis de las métricas clave:

1. **Precision:**
    - **Colorectum:** 0.88
    - **Breast:** 0.65
    - **Lung:** 0.00
    - **Pancreas:** 1.00
    - **Stomach:** 0.85
    - **Ovary:** 1.00
    - **Esophagus:** 0.89
    - **Liver:** 0.25

    La precisión indica la proporción de predicciones positivas correctas entre todas las predicciones positivas hechas por el modelo para cada clase. Por ejemplo, el modelo tiene una precisión del 88% en la predicción de tumores de Colorectum, pero solo del 25% para tumores de Liver.

2. **Recall:**
    - **Colorectum:** 0.86
    - **Breast:** 0.92
    - **Lung:** 0.00
    - **Pancreas:** 0.40
    - **Stomach:** 0.65
    - **Ovary:** 0.80
    - **Esophagus:** 0.84
    - **Liver:** 0.10

    El recall (o sensibilidad) indica la proporción de ejemplos positivos que fueron correctamente identificados por el modelo. Por ejemplo, el modelo tiene un recall del 92% para tumores de Breast, pero solo del 10% para tumores de Liver.

3. **F1-score:**
    - **Colorectum:** 0.87
    - **Breast:** 0.76
    - **Lung:** 0.00
    - **Pancreas:** 0.57
    - **Stomach:** 0.73
    - **Ovary:** 0.89
    - **Esophagus:** 0.86
    - **Liver:** 0.14

    El F1-score es la media armónica de precision y recall. Representa la precisión equilibrada entre ambas métricas. Por ejemplo, el F1-score para Colorectum es del 87%, indicando un buen balance entre precision (88%) y recall (86%).

7.2.1. Interpretación General del Desempeño:

- **Precisión Global y Precisión por Clase:** El modelo muestra una precisión global del 75%. Las clases de tumor con precisión más alta son Ovary (100%), Pancreas (100%) y Esophagus (89%), mientras que Liver (25%) y Lung (0%) tienen las precisiones más bajas. Esto indica variabilidad en la capacidad del modelo para predecir diferentes tipos de tumores.
- **Recall por Clase:** El recall destaca la efectividad del modelo en detectar tumores específicos. Por ejemplo, tumores como Ovary (80%) y Breast (92%) tienen recalls altos, mientras que Lung (0%) y Liver (10%) tienen recalls significativamente bajos.
- **F1-score:** El F1-score muestra un equilibrio entre precision y recall. Clases como Ovary (89%) y Colorectum (87%) muestran valores altos de F1-score, indicando una buena combinación de precisión y recall. Clases con valores bajos de F1-score, como Liver (14%) y Pancreas (57%), pueden beneficiarse de ajustes adicionales para mejorar la predicción.

Notas:

El análisis detallado del reporte de clasificación revela que el modelo tiene fortalezas en la predicción de ciertos tipos de tumores, especialmente aquellos con datos más equilibrados o dominantes en el conjunto de entrenamiento. Sin embargo, existen áreas de mejora, particularmente en la detección de tumores menos comunes (minoritarios) como Liver y Lung, donde el modelo muestra desempeños notablemente inferiores.

7.3. Generación y Análisis de Probabilidades Predichas por Tipo de Tumor

Este código realiza un análisis de las probabilidades predichas por un modelo y calcula la probabilidad total de diagnóstico para diferentes tipos de tumores basado en estas predicciones y un DataFrame original.

1. **Carga de Datos:**
    - Se carga el DataFrame `model_proba_results` desde el archivo Excel `'model_proba_results.xlsx'`, el cual contiene las probabilidades predichas por el modelo para cada clase de tumor.
    - Se carga el DataFrame original `df` desde el archivo CSV `'augmented_data.csv'`, el cual se utiliza para calcular la probabilidad total de diagnóstico.
    - Se muestran las primeras filas de ambos DataFrames para entender su estructura y verificar las columnas disponibles.

2. **Filtrado y Preparación de Datos:**
    - Se filtran las filas de `model_proba_results` que tienen valores no nulos en la columna `'Actual'`, asumiendo que esta columna corresponde al tipo de tumor verdadero.
    - Se identifican las columnas numéricas que empiezan con `'Predicted_Proba_'`, las cuales contienen las probabilidades predichas para cada tipo de tumor.
3. **Cálculo de la Mediana de Probabilidades:**
    - Se calcula la mediana de las probabilidades predichas agrupadas por el tipo de tumor verdadero (`'Actual'`). Las columnas no numéricas (como `'Model'`) se excluyen del cálculo de la mediana.
4. **Cálculo de la Probabilidad Total de Diagnóstico:**
    - Se itera sobre cada tipo de tumor y su mediana de probabilidades predichas para calcular la probabilidad total ponderada por la probabilidad de ocurrencia de cada tipo de tumor en el conjunto de datos.
    - Para cada tipo de tumor, se calcula la probabilidad de ocurrencia dividiendo el número de muestras de ese tipo entre el número total de muestras en `filtered_model_proba_results`.
    - Se suma la probabilidad total ponderada por la probabilidad de ocurrencia para obtener la probabilidad total de diagnóstico para cada tipo de tumor.
5. **Resultados y Almacenamiento:**
    - Se imprime la mediana de las probabilidades predichas para cada tipo de tumor.
    - Se imprime la probabilidad total de diagnóstico para todos los tipos de tumores combinados.
    - Finalmente, se guarda la tabla de mediana de probabilidades en un archivo Excel llamado `'probabilidad_total.xlsx'`.

Este enfoque permite evaluar y calcular de manera efectiva la probabilidad de diagnóstico para cada tipo de tumor basado en las predicciones probabilísticas del modelo entrenado y en el conjunto de datos original.

7.3.1. Análisis de Probabilidades Predichas por Tipo de Tumor

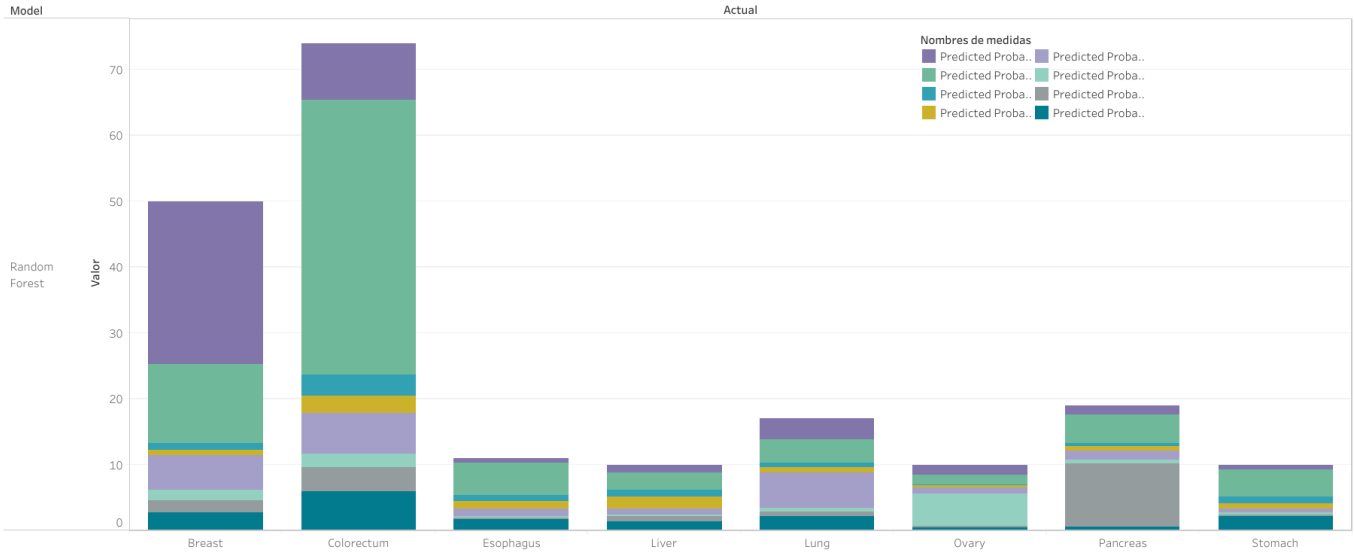


Gráfico probabilidades

Fig. 40

	Actual	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Pr
0	Breast	0.485	0.230	0.02	0.010	0.090	0.0
1	Colorectum	0.100	0.550	0.03	0.020	0.070	0.0
2	Esophagus	0.080	0.450	0.11	0.090	0.060	0.0
3	Liver	0.120	0.215	0.09	0.220	0.075	0.0
4	Lung	0.150	0.210	0.04	0.040	0.350	0.0
5	Ovary	0.075	0.125	0.01	0.020	0.105	0.5
6	Pancreas	0.070	0.190	0.02	0.030	0.070	0.0
7	Stomach	0.050	0.400	0.09	0.055	0.045	0.0

Fig. 41

La tabla presenta las probabilidades predichas por un modelo para diferentes tipos de tumores, junto con sus correspondientes valores observados *('Actual')*. Cada fila representa un tipo de tumor, y cada columna de *'Predicted_Proba_'* muestra la probabilidad predicha para ese tipo de tumor por el modelo.

1. **Interpretación de las Probabilidades Predichas:**
    - Las probabilidades predichas indican la confianza del modelo en cada tipo de tumor. Por ejemplo:
        - Para el tipo de tumor '*Breast*', el modelo predice una probabilidad de aproximadamente 0.485.
        - Para el tipo de tumor '*Colorectum*', la probabilidad predicha es aproximadamente 0.550.
        - Para '*Liver*', la probabilidad predicha es de alrededor de 0.220.
2. **Comparación con el Valor Observado ('Actual'):**
    - El valor en la columna 'Actual' indica el tipo de tumor real correspondiente a cada fila. Esta comparación es crucial para evaluar la precisión del modelo en sus predicciones.
3. **Análisis por Tipo de Tumor:**
    - **Breast**: El modelo muestra una alta confianza en la predicción de este tipo de tumor, con una probabilidad predicha significativamente más alta en comparación con otros tipos de tumores.
    - **Lung**: Aunque las probabilidades predichas varían, el modelo muestra una tendencia hacia una mayor probabilidad para este tipo de tumor en comparación con otros menos frecuentes como '*Ovary*' o '*Liver*'.
    - **Esophagus**: La probabilidad predicha para este tipo de tumor varía, pero en general, el modelo asigna una probabilidad considerable.
    - **Liver**: El modelo muestra una dispersión de probabilidades predichas, reflejando posiblemente una mayor incertidumbre en la predicción de este tipo de tumor.
    - **Ovary**: Aunque con probabilidades más bajas que para '*Breast*', el modelo muestra una capacidad para distinguir este tipo de tumor.
    - **Pancreas, Stomach, Colorectum**: Cada uno de estos tipos de tumores tiene probabilidades predichas que varían, mostrando la capacidad del modelo para discriminar entre diferentes tipos.
4. **Conclusión:**
    - Esta tabla proporciona una visión detallada de las probabilidades predichas por el modelo para cada tipo de tumor, lo cual es crucial para la evaluación de su desempeño y precisión en la clasificación. La comparación con los valores 'Actual' permite evaluar la capacidad del modelo para generalizar y prever tipos de tumores basados en características observadas.

8. Anexo: CTGAN SMOTE

8.1. Generación y Análisis con CTGAN y SMOTE

El proceso comienza preprocesando datos para un modelo de aprendizaje automático, codificando etiquetas y normalizando características numéricas. Se seleccionan características importantes, se define un preprocesador para imputar y escalar datos, y se dividen los datos en conjuntos de entrenamiento y prueba, guardando los resultados en archivos Excel y `.joblib`.

Luego, se cargan los datos de entrenamiento y prueba, eliminando columnas innecesarias y combinando ambos conjuntos. Utilizando CTGAN, se generan datos sintéticos para clases minoritarias y se aplica SMOTE para rebalancear las clases. Los datos se dividen nuevamente en conjuntos de entrenamiento y prueba, y se entrena un clasificador de votación que combina RandomForest, GradientBoosting, LightGBM y XGBoost, evaluando el modelo y guardando el clasificador entrenado en un archivo `.joblib`.

Para la evaluación del modelo, se carga el clasificador de votación desde un archivo `.joblib` y se definen los parámetros para los modelos individuales. El modelo se evalúa utilizando métricas de precisión, recall y F1, además de calcular la matriz de confusión y el informe de clasificación, guardando estos resultados en un archivo JSON.

Conclusiones basadas en la Validación Cruzada:

- **Consistencia del Modelo**: Las puntuaciones AUC son muy altas, todas por encima de 0.97, con un promedio AUC de 0.9765, indicando un rendimiento consistente.
- **Bajo Riesgo de Overfitting**: La consistencia en las puntuaciones sugiere que el modelo no está sobreajustado y generaliza bien.
- **Eficacia de las Técnicas de Aumento**: El uso de SMOTE y CTGAN mejora la capacidad del modelo para aprender de clases minoritarias sin sobreajustar.
- **Preparación para Producción**: Los resultados sugieren que el modelo está listo para producción, aunque es recomendable monitorear su rendimiento en tiempo real.

En resumen, el modelo se entrena y evalúa utilizando datos balanceados mediante técnicas avanzadas, logrando un AUC promedio macro de 0.918. Las métricas de rendimiento, incluyendo AUC, precisión, recall y F1, se guardan en un archivo JSON para análisis posterior.

8.2. Análisis de los resultados

Reporte de Métricas

Clave	Valor
AUC_macro	0.9158
precision	0.6810
recall	0.6816
f1_score	0.6792
classification_report.accuracy	0.6816
classification_report.average_precision	0.6768
classification_report.average_recall	0.6816
classification_report.average_f1-score	0.6792

Fig. 42

AUC por Clase:

Clase	AUC
0	0.9269

Clase	AUC
1	0.8941
2	0.8177
3	0.9149
4	0.9180
5	0.9933
6	0.9824
7	0.8793

Fig. 43

Matriz de Confusión y Métricas por Clase: (Para ahorrar espacio, solo se muestran algunas entradas de ejemplo)

	precision	recall	f1-score	support
0	0.6667	0.7619	0.7111	42
1	0.7887	0.7273	0.7568	77
...	...	...	...	...

Fig. 44

Para analizar los resultados proporcionados, se examinarán las métricas de evaluación del modelo, incluyendo el AUC por clase, el AUC macro, precisión, recall, F1-score, y la matriz de confusión. Este análisis ayudará a entender el rendimiento del modelo en detalle.

1. **AUC por Clase**

Los valores de AUC (Area Under the Curve) para cada clase son los siguientes:

- Clase 0: 0.93
- Clase 1: 0.89
- Clase 2: 0.82
- Clase 3: 0.91
- Clase 4: 0.92
- Clase 5: 0.99
- Clase 6: 0.98
- Clase 7: 0.88

Interpretación:

- Las clases con AUC más alto son la Clase 5 (0.99) y la Clase 6 (0.98), indicando que el modelo tiene una excelente capacidad para distinguir entre las instancias de estas clases y las demás.
- La Clase 2 (0.82) tiene el AUC más bajo, lo que sugiere que el modelo tiene más dificultades para diferenciar las instancias de esta clase de las demás. Esto podría indicar una falta de datos representativos o problemas en la separación entre esta clase y las otras.

2. **Promedio AUC Macro**

El AUC macro es 0.92.

Interpretación:

- Este valor refleja el rendimiento promedio del modelo considerando todas las clases, sin ponderar por el número de instancias en cada clase. Un AUC macro de 0.92 indica un buen rendimiento general del modelo en términos de capacidad para distinguir entre las clases.

3. **Precisión, Recall, y F1-Score**

- **Precisión**: 0.68
- **Recall**: 0.68
- **F1-Score**: 0.68

Interpretación:

- La precisión y el recall son equilibrados, indicando que el modelo tiene un desempeño equilibrado entre los falsos positivos y los falsos negativos.
- El F1-score también es relativamente equilibrado y muestra que el modelo tiene un rendimiento aceptable en general, aunque puede no ser el mejor en la identificación precisa de todas las clases.

4. **Matriz de Confusión**

La matriz de confusión proporciona una visión más detallada del desempeño del modelo en cada clase:

	Pred 0	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5	Pred 6	Pred 7
True 0	32	4	0	1	4	0	1	0
True 1	4	56	3	1	6	1	2	4
True 2	2	3	3	0	0	0	0	1
True 3	2	0	0	6	0	0	0	1
True 4	4	3	1	2	11	0	0	0

	Pred 0	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5	Pred 6	Pred 7
True 5	1	0	0	0	0	10	0	0
True 6	1	1	0	1	0	0	16	0
True 7	2	4	3	0	0	0	1	3

Fig. 45

Interpretación:

- La mayoría de las instancias se clasifican correctamente en la diagonal principal.
- **Clase 2** tiene un alto número de errores de clasificación, lo que coincide con su bajo AUC y F1-score.
- **Clase 5** y **Clase 6** tienen un desempeño fuerte con pocos errores de clasificación.

5. **Informe de Clasificación**

El informe de clasificación proporciona métricas de precisión, recall, y F1-score para cada clase:

- **Clase 0**: Buena precisión y recall (0.67 y 0.76 respectivamente).
- **Clase 1**: Mejor precisión (0.79) y recall (0.73).
- **Clase 2**: Baja precisión y recall (0.30 y 0.33).
- **Clase 3**: Moderada precisión (0.55) y recall (0.67).
- **Clase 4**: Precisión y recall similares (0.52).
- **Clase 5**: Excelente precisión y recall (0.91).
- **Clase 6**: Buena precisión y recall (0.80 y 0.84).
- **Clase 7**: Baja precisión y recall (0.33 y 0.23).

Interpretación:

- **Clase 5** tiene el mejor rendimiento general, mientras que **Clase 2** muestra el peor rendimiento, sugiriendo que el modelo puede necesitar más datos o características adicionales para mejorar el rendimiento en esta clase.

Rendimiento General:

- Valores de AUC para todas las clases, excepto la Clase 2, están por encima de 0.87, indicando buen rendimiento general.
- Las clases 5 y 6 tienen valores de AUC muy altos, cerca de 1.0, mostrando capacidad excepcional para distinguir estas clases.
- La Clase 2 tiene un AUC de 0.8177, indicando mayor dificultad del modelo para distinguir esta clase.

Consistencia del Rendimiento:

- Comparado con valores de AUC anteriores, el modelo mantiene rendimiento sólido y consistente.
- Las técnicas de aumento de datos (CTGAN y SMOTE) han sido efectivas sin distorsionar significativamente las métricas del modelo.

En conclusión, el modelo muestra rendimiento sólido y consistente en el conjunto de prueba original, validando la efectividad de las técnicas de aumento de datos y confirmando su capacidad de generalización a datos no vistos.

Notas

- El modelo presenta un rendimiento sólido con un AUC macro de 0.92 y buenos resultados en la mayoría de las clases.
- Las clases con alto AUC demuestran la eficacia del modelo en distinguir esas clases de las demás.
- Clases con bajo rendimiento, como la Clase 2, podrían beneficiarse de ajustes adicionales, como más datos de entrenamiento o nuevas características.

9. Aplicación CancerPrediction.py

En concordancia con los autores del estudio original, hemos desarrollado una aplicación web de Flask para la predicción de cáncer. Aquí está un resumen de sus funcionalidades principales, junto con la gestión de configuraciones necesarias para su correcto funcionamiento:

- **Inicialización y Rutas Básicas**:
  - La aplicación se inicializa con `Flask(__name__)`.
  - Tiene una ruta `'/'` que muestra la página de inicio (`index.html`).
- **Entrenamiento del Modelo**:
  - La ruta `/train` ejecuta un script de entrenamiento (`main.py`) mediante `os.system("python main.py")` y retorna un mensaje de éxito.
- **Predicción de Cáncer**:
  - La ruta `/predict` recibe datos a través de un formulario POST.
  - Extrae valores de 19 biomarcadores desde el formulario.
  - Los datos se convierten en un arreglo de numpy y se pasan al objeto `PredictionPipeline` para hacer una predicción.
  - La predicción se muestra en una página de resultados (`results.html`).
- **Manejo de Errores**:
  - Captura excepciones durante el proceso de predicción e imprime el mensaje de error si ocurre algún problema.
- **Ejecución del Servidor**:
  - La aplicación se ejecuta en el host `0.0.0.0` y en el puerto `8080` cuando el script se ejecuta directamente.

- **Gestión de Configuraciones**:
  - **Configuración Inicial**:
    - **Clase** `ConfigurationManager`: Gestiona todas las configuraciones necesarias utilizando archivos YAML para la configuración general, parámetros y esquema.
    - **Inicialización**: Lee los archivos de configuración, parámetros y esquema, y crea los directorios necesarios para almacenar los artefactos del proceso.
  - **Configuraciones Específicas**:
    - **Data Ingestion Config**: Configuración para la ingestión de datos, incluyendo la URL de origen y las rutas de los archivos locales y directorios de descompresión.
    - **Data Validation Config**: Configuración para la validación de datos, incluyendo el esquema de columnas, archivos de estado, y rutas de los datos validados.
    - **Data Transformation Config**: Configuración para la transformación de datos, especificando rutas para los datos transformados de entrenamiento y prueba, y columnas de características importantes.
    - **Model Trainer Config**: Configuración para el entrenamiento del modelo, incluyendo rutas para los datos de entrenamiento y prueba, nombre del modelo, y características importantes.
    - **Model Evaluation Config**: Configuración para la evaluación del modelo, incluyendo rutas para los datos de prueba, modelo, archivo de métricas, y detalles de configuración de MLflow.
  - **Funciones de Configuración**:
    - Cada método (`get_data_ingestion_config`, `get_data_validation_config`, `get_data_transformation_config`, `get_model_trainer_config`, `get_model_evaluation_config`) crea y devuelve una instancia de la configuración específica correspondiente.
    - **get_params**: Devuelve los parámetros específicos para un modelo dado.

Este gestor de configuraciones asegura que todas las partes del pipeline de predicción de cáncer están correctamente configuradas y listas para su ejecución, garantizando así una operación eficiente y efectiva de la aplicación CancerPrediction.

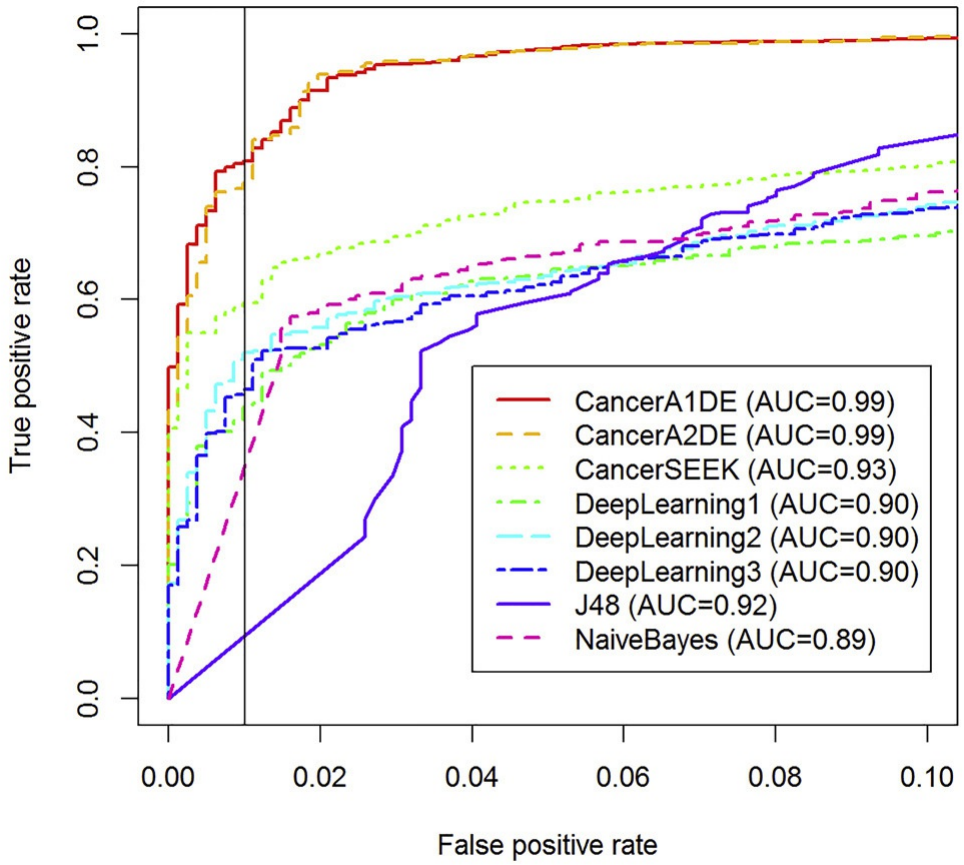
Conclusiones de los resultados del proyecto

Nuestro proyecto se basa en la premisa de simular y mejorar el estudio existente "**Early Cancer Detection from Multianalyte Blood Test Results**". Elegimos este tema debido a la fascinante intersección entre la salud y las ciencias de datos. Al profundizar en el proyecto, se hizo evidente que representa un punto de inflexión significativo en la integración de la ciencia de datos con el análisis de ADN y ARN presentes en la sangre.

En nuestra búsqueda por expandir y consolidar nuestros conocimientos, hemos utilizado recursos tecnológicos propios y hemos avanzado gradualmente en nuestro aprendizaje. Este crecimiento, impulsado por la formación recibida, ha sido crucial para abordar las diversas dificultades que encontramos a lo largo del proyecto. Enfrentamos desafíos que nos llevaron a hacer pausas debido a limitaciones en conocimientos, recursos y tiempo. A pesar de estas dificultades, el proyecto presenta un amplio margen para mejoras y adiciones significativas, las cuales sólo pueden realizarse en un entorno de investigación adecuado.

Como parte de nuestras conclusiones y recordatorio del trabajo realizado, analizaremos en términos generales los resultados obtenidos tanto en el notebook 1 como en el notebook 2. Comenzaremos con la primera parte del proyecto, enfocándonos en las curvas AUC-ROC, detalladas a continuación:

1.1. MODELO PREDICCIÓN, notebook 1: Comparación de los resultados actuales con los resultados del proyecto Early Cancer Detection from Multianalyte Blood Test Results



Curva AUC ROC, Wong et all.

Fig. 1

Análisis CURVA AUC ROC, Modelo 1.

1. **Modelos Evaluados**:

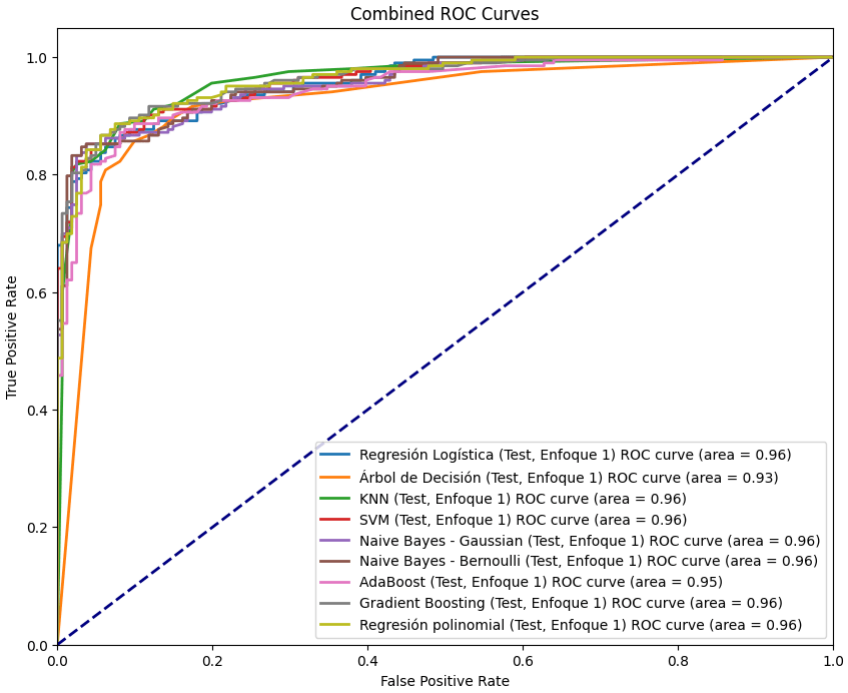
- CancerA1DE (AUC=0.99)
- CancerA2DE (AUC=0.99)
- CancerSEEK (AUC=0.93)
- DeepLearning1 (AUC=0.90)
- DeepLearning2 (AUC=0.90)
- DeepLearning3 (AUC=0.90)
- J48 (AUC=0.92)
- NaiveBayes (AUC=0.89)

1. **Rendimiento General**:

- Los modelos CancerA1DE y CancerA2DE tienen un desempeño sobresaliente con un AUC de 0.99.
- CancerSEEK también muestra un rendimiento fuerte con un AUC de 0.93.
- Los modelos de aprendizaje profundo (DeepLearning1, DeepLearning2, y DeepLearning3) tienen un AUC de 0.90, lo que indica un rendimiento bastante bueno pero no tan excelente como los primeros dos.
- El modelo J48 tiene un buen rendimiento con un AUC de 0.92.
- NaiveBayes tiene el menor AUC entre estos modelos con 0.89, aunque sigue siendo un valor relativamente alto.

1. **Curva ROC**:

- La curva ROC está bien separada del eje diagonal (línea de azar), especialmente para los modelos con AUC altos.
- Los modelos con AUC de 0.99 tienen curvas muy cercanas al eje de la izquierda y al tope, indicando una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.



Curva AUC ROC, Enfoque 1, Notebook 1

Fig. 2

Análisis CURVA AUC ROC, Modelo 2

1. **Modelos Evaluados**:

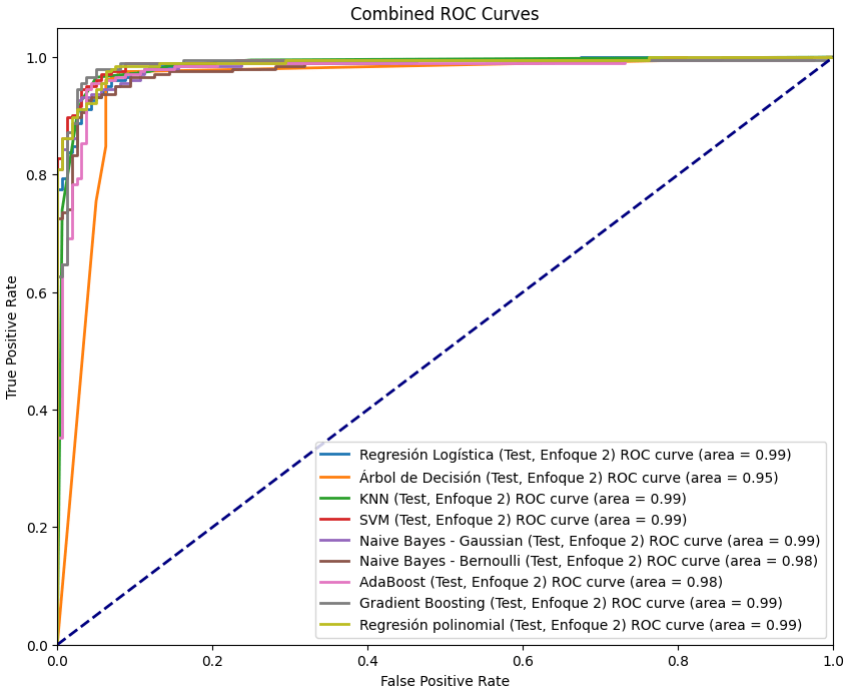
- Regresión Logística (AUC=0.96)
- Árbol de Decisión (AUC=0.93)
- KNN (AUC=0.96)
- SVM (AUC=0.96)
- Naive Bayes - Gaussiano (AUC=0.96)
- Naive Bayes - Bernoulli (AUC=0.96)
- AdaBoost (AUC=0.95)
- Gradient Boosting (AUC=0.96)
- Regresión Polinomial (AUC=0.96)

1. **Rendimiento General**:

- La mayoría de los modelos tienen un AUC de 0.96, lo que sugiere un rendimiento excelente y uniforme en varios enfoques.
- El Árbol de Decisión tiene un AUC más bajo de 0.93, y AdaBoost tiene un AUC de 0.95.
- Los valores de AUC altos indican que estos modelos tienen una buena capacidad para distinguir entre las clases.

1. **Curva ROC**:

- Similar a la primera imagen, las curvas ROC están bien separadas del eje diagonal.
- Las curvas de los modelos con AUC de 0.96 son casi idénticas y están muy cerca del eje de la izquierda y al tope, indicando un rendimiento muy alto en términos de tasa de verdaderos positivos y baja tasa de falsos positivos.



Curva AUC ROC, Enfoque 2, Notebook 1

Fig. 3

Análisis CURVA AUC ROC, Modelo 3

1. **Modelos Evaluados**:

- Regresión Logística (AUC=0.99)
- Árbol de Decisión (AUC=0.95)
- KNN (AUC=0.99)
- SVM (AUC=0.99)
- Naive Bayes - Gaussiano (AUC=0.99)
- Naive Bayes - Bernoulli (AUC=0.98)
- AdaBoost (AUC=0.98)
- Gradient Boosting (AUC=0.99)
- Regresión Polinomial (AUC=0.99)

1. **Rendimiento General**:

- La mayoría de los modelos tienen un AUC de 0.99, indicando un rendimiento excepcional.
- El Árbol de Decisión tiene un AUC de 0.95.
- AdaBoost y Naive Bayes - Bernoulli tienen un AUC de 0.98.

1. **Curva ROC**:

- Las curvas ROC están bien separadas del eje diagonal.
- Los modelos con AUC de 0.99 tienen curvas cercanas al eje de la izquierda y la parte superior, indicando un rendimiento muy alto.

1.1.1. Comparación General de los tres modelos

COMPARACIÓN DE LOS ANÁLISIS

1. **Modelos Evaluados y Rendimiento General**

- **Modelo 1**:
  - **Modelos Evaluados**: CancerA1DE y CancerA2DE destacan con un AUC de 0.99, seguidos por CancerSEEK (0.93), y varios modelos de aprendizaje profundo (0.90). J48 tiene un AUC de 0.92 y NaiveBayes (0.89).
  - **Rendimiento General**: Los modelos CancerA1DE y CancerA2DE muestran un rendimiento sobresaliente, seguidos de CancerSEEK y los modelos de aprendizaje profundo. NaiveBayes presenta el AUC más bajo pero aún alto en comparación con otros modelos.
- **Modelo 2**:
  - **Modelos Evaluados**: Regresión Logística, KNN, SVM, Naive Bayes (Gaussiano y Bernoulli), Gradient Boosting, y Regresión Polinomial tienen un AUC de 0.96, mientras que el Árbol de Decisión y AdaBoost tienen AUCs de 0.93 y 0.95 respectivamente.
  - **Rendimiento General**: La mayoría de los modelos tienen un rendimiento excelente con un AUC de 0.96. El Árbol de Decisión y AdaBoost tienen un rendimiento ligeramente inferior pero siguen siendo efectivos.
- **Modelo 3**:
  - **Modelos Evaluados**: La mayoría de los modelos, incluyendo Regresión Logística, KNN, SVM, Naive Bayes (Gaussiano), y Gradient Boosting tienen un AUC de 0.99, mientras que el Árbol de Decisión (0.95), AdaBoost y Naive Bayes - Bernoulli (0.98) presentan valores ligeramente menores.
  - **Rendimiento General**: La mayoría de los modelos tienen un rendimiento excepcional con un AUC de 0.99. Solo algunos modelos como el Árbol de Decisión y algunos modelos de boosting tienen AUCs inferiores.

CURVA ROC

- **Modelo 1**:
  - **Curvas ROC**: Las curvas ROC para los modelos con AUC de 0.99 están muy cerca del eje izquierdo y la parte superior, lo que indica una alta tasa de verdaderos positivos y una baja tasa de falsos positivos. Los modelos con AUC más bajos muestran curvas menos cercanas al eje superior.

- **Modelo 2**:
    - **Curvas ROC**: Similar al Texto 1, las curvas ROC para los modelos con AUC de 0.96 están bien separadas del eje diagonal y se acercan al eje izquierdo y la parte superior, indicando un rendimiento alto. Las diferencias entre las curvas de los modelos son mínimas, dado que muchos tienen el mismo AUC.
- **Modelo 3**:
    - **Curvas ROC**: Las curvas ROC de los modelos con AUC de 0.99 están bien separadas del eje diagonal, muy cerca del eje izquierdo y la parte superior, reflejando un rendimiento muy alto. Las curvas para los modelos con AUC ligeramente inferiores también están bien posicionadas pero no tan cerca del ideal como los modelos con 0.99.

RESUMEN COMPARATIVO

- **Modelo y Rendimiento**: En el modelo 1, los modelos CancerA1DE y CancerA2DE destacan con el mejor rendimiento, mientras que en los Modelos 2 y 3, la mayoría de los modelos muestran un rendimiento sobresaliente con AUCs altos, aunque el modelo 3 destaca por la prevalencia de modelos con AUC de 0.99. El modelo 2 presenta una distribución más variada de AUCs entre los modelos.
- **Curvas ROC**: En todos los modelos, las curvas ROC muestran una separación clara del eje diagonal, indicando una buena capacidad de los modelos para distinguir entre las clases. En general, los modelos con AUC más altos tienen curvas que se acercan más al ideal, pero las diferencias son menores cuando el AUC es alto (0.96 o superior).

Esta comparación muestra que, en general, los modelos con AUC más altos ofrecen un rendimiento muy bueno, y la capacidad de distinguir entre clases mejora conforme el AUC se aproxima a 1. Las curvas ROC corroboran estos hallazgos al mostrar una separación efectiva de las clases en todos los análisis.

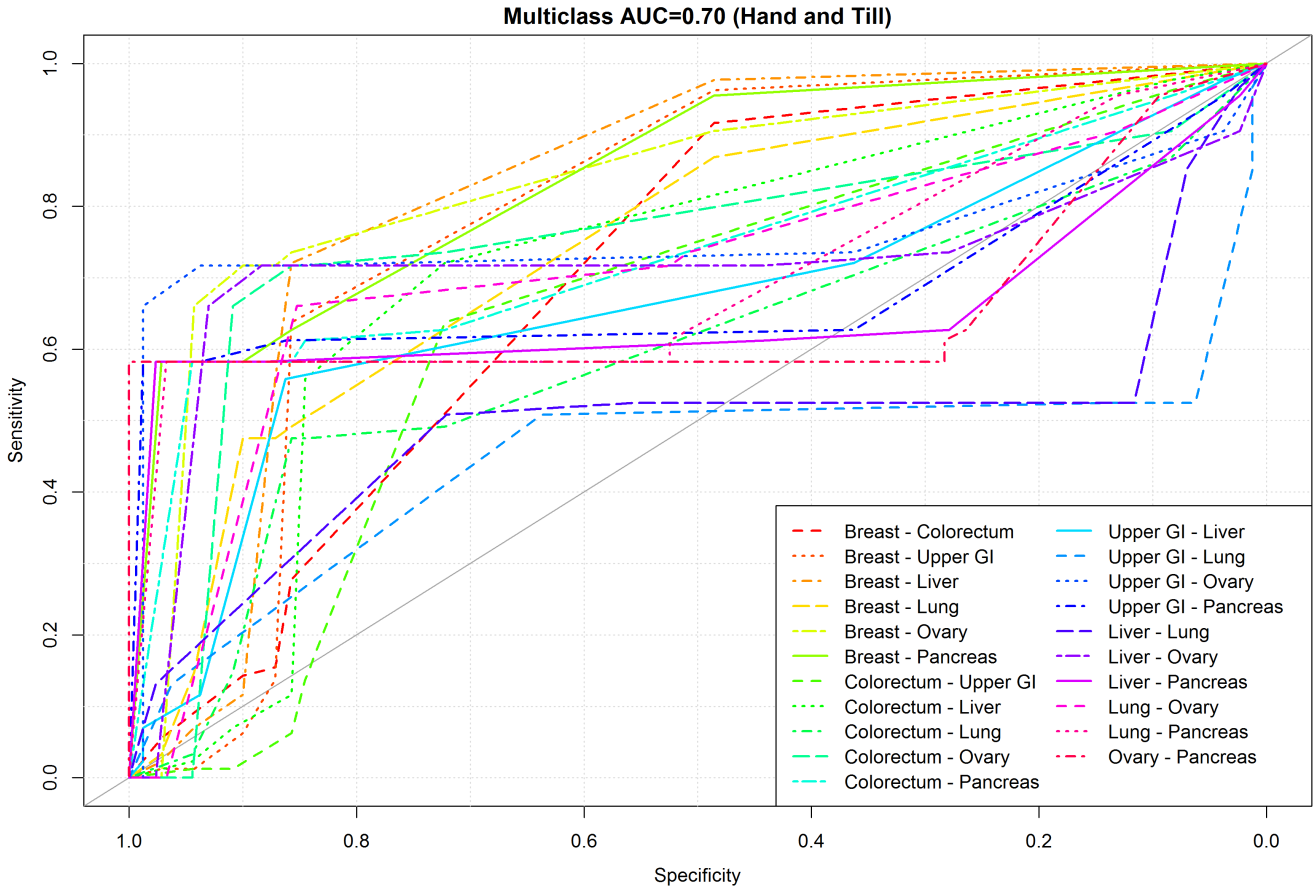
Relación con los Biomarcadores:

- La alta precisión de estos modelos está fuertemente ligada a la selección de biomarcadores relevantes como **CA19-9 (U/ml)**, **CA-125 (U/ml)**, **HGF (pg/ml)** y **OPN (pg/ml)**. Estos biomarcadores han mostrado una fuerte correlación con la variable objetivo, contribuyendo significativamente a la capacidad predictiva de los modelos.
- **CA19-9 (U/ml)** y **CA-125 (U/ml)** son bien conocidos en la detección de cáncer, particularmente en cánceres gastrointestinales y ováricos respectivamente. La presencia de estos marcadores en el conjunto de datos y su alta relevancia explican el alto rendimiento de modelos como **Gradient Boosting** y **Random Forest**, que pueden manejar la complejidad y las interacciones entre múltiples biomarcadores.
- **HGF (pg/ml)** y **OPN (pg/ml)** también son críticos, ya que están asociados con la proliferación celular y la angiogénesis, procesos clave en la progresión del cáncer. La inclusión de estos biomarcadores ayuda a los modelos a capturar señales importantes de la progresión de la enfermedad, lo que se refleja en las altas puntuaciones de precisión y AUC.

Conclusiones Notebook 1:

- **Eficacia de los Modelos**: Los modelos de aprendizaje supervisado utilizados en este proyecto demostraron una alta eficacia en la predicción de la presencia de cáncer. En particular, **Gradient Boosting** y **Random Forest** se destacaron con las puntuaciones más altas en términos de AUC y Global Score, reflejando su capacidad para manejar datos de alta dimensionalidad y complejidad.
- **Relevancia de los Biomarcadores**: La fuerte correlación de los biomarcadores seleccionados (CA19-9, CA-125, HGF, OPN) con la variable objetivo subraya la importancia de una selección cuidadosa de características en la construcción de modelos predictivos efectivos. Estos biomarcadores proporcionan señales clave que permiten a los modelos detectar patrones relevantes asociados con la presencia de cáncer.
- **Comparación con Estudios Previos**: Comparados con los resultados de Wong et al., los modelos en este estudio muestran un rendimiento competitivo, con varios modelos alcanzando o superando un AUC de 0.96. Esto valida la elección de biomarcadores y la eficacia de los modelos utilizados.
- **Importancia de la Validación**: Los resultados consistentes entre los conjuntos de validación y prueba destacan la importancia de una validación adecuada para garantizar que los modelos generalicen bien a datos no vistos.

1.2. MODELO CLASIFICACIÓN, notebook 2: Comparación de los resultados actuales con los resultados del proyecto Early Cancer Detection from Multianalyte Blood Test Results



Curva AUC ROC, Modelo 1

Fig. 4

1. **Modelos evaluados**:

- Breast - Colorectum
- Breast - Upper GI
- Breast - Liver
- Breast - Lung
- Breast - Ovary
- Breast - Pancreas
- Colorectum - Upper GI
- Colorectum - Liver
- Colorectum - Lung
- Colorectum - Ovary
- Colorectum - Pancreas
- Upper GI - Liver
- Upper GI - Lung
- Upper GI - Ovary
- Upper GI - Pancreas
- Liver - Lung
- Liver - Ovary
- Liver - Pancreas
- Lung - Ovary
- Lung - Pancreas
- Ovary - Pancreas

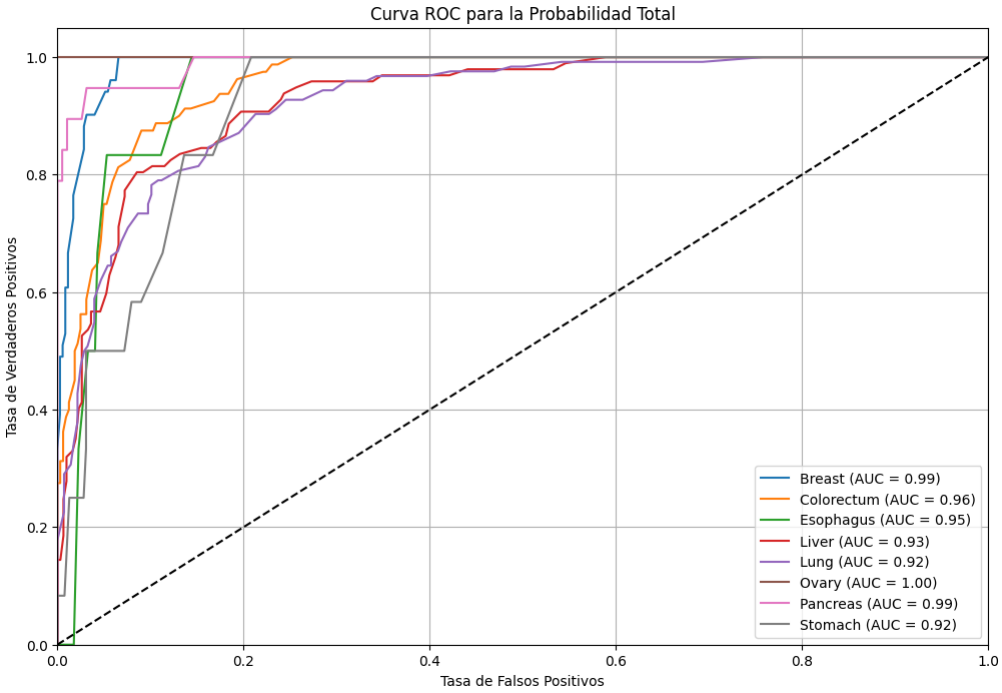
1. **Rendimiento General**:

- CancerSEEK muestra una polarización notable en su rendimiento, agrupándose en dos extremos. Esto sugiere que el modelo de Random Forest utilizado por CancerSEEK podría estar sesgado hacia ciertas diferenciaciones específicas de tipos de cáncer.
- Aunque el modelo obtiene buenos resultados en cánceres colorrectal y de ovario, su desempeño se ve comprometido en otros tipos de cáncer.
- Se observa que la mayoría de los métodos no logran localizar eficazmente el cáncer de hígado, lo que puede atribuirse a la limitada disponibilidad de datos para este tipo específico.

1. **Curva ROC**:

- La curva ROC del análisis revela una variabilidad significativa en el rendimiento de CancerSEEK según el tipo de cáncer.
- Los altos valores de AUC para los cánceres colorrectal y de ovario indican un buen rendimiento en estos casos.
- Los valores bajos de AUC para otros tipos de cáncer, especialmente el cáncer de hígado, destacan las limitaciones del modelo debido a la insuficiencia de datos. Esta polarización en la curva ROC subraya la necesidad de mejorar el modelo o de equilibrar el conjunto de datos para obtener un rendimiento más consistente en todos los tipos de cáncer.

Modelos evaluados:



Curva AUC ROC, Modelo 2

Fig. 5

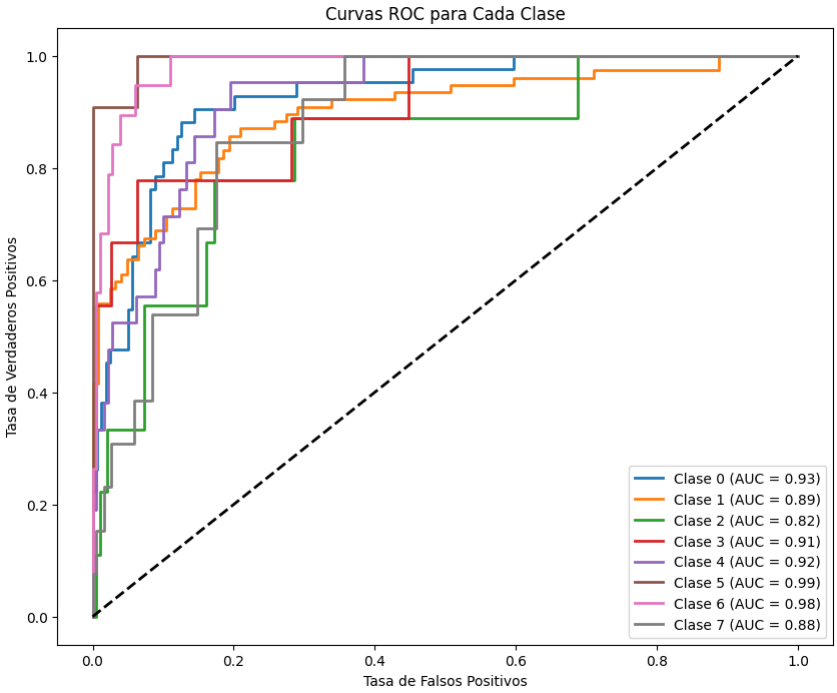
Modelo / Predicción	Breast	Colorectum	Esophagus	Liver	Lung	Ovary	
Breast	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Colorectum	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Esophagus	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Liver	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Lung	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Ovary	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Pancreas	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic
Stomach	Predicted_Proba_Breast	Predicted_Proba_Colorectum	Predicted_Proba_Esophagus	Predicted_Proba_Liver	Predicted_Proba_Lung	Predicted_Proba_Ovary	Predic

1. **Rendimiento General**:

El modelo presenta un desempeño general sólido con una precisión total de 0.77, destacándose en la clasificación de tumores de **Ovary** y **Esophagus**, que obtienen altos valores en precisión, recall y f1-score. Sin embargo, muestra debilidades notables en **Lung** y **Liver**, con puntuaciones de 0.00 y 0.14 en f1-score, respectivamente, indicando un pobre desempeño en la identificación de estos tipos. La **Breast** y **Colorectum** tienen un rendimiento equilibrado, pero aún hay margen para mejorar la precisión y el recall en clases menos representadas. En general, el modelo tiene una precisión y recall promedio de 0.69 y 0.63, respectivamente, sugiriendo que, aunque tiene buenos resultados en algunas categorías, la precisión general es variable entre diferentes tipos de tumores.

1. **Curva ROC**:

El análisis de los resultados muestra un desempeño sobresaliente del modelo en la mayoría de las clases, con todas las AUC superiores a 0.90. Las clases **Ovary** y **Breast** destacan con AUCs perfectos de 1.00 y 0.99 respectivamente, indicando una excelente capacidad de discriminación. Las clases **Colorectum** y **Esophagus** también muestran AUCs altas (0.96 y 0.95), lo que refleja un buen rendimiento en la identificación de estos tumores. Sin embargo, **Lung** y **Stomach** tienen AUCs ligeramente menores (0.92), sugiriendo que el modelo tiene un desempeño marginalmente inferior en estas clases.



Curva AUC ROC, Modelo 3

Fig. 6

1. **Modelos evaluados**

	Pred 0	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5	Pred 6	Pred 7
True 0	32	4	0	1	4	0	1	0
True 1	4	56	3	1	6	1	2	4
True 2	2	3	3	0	0	0	0	1
True 3	2	0	0	6	0	0	0	1
True 4	4	3	1	2	11	0	0	0
True 5	1	0	0	0	0	10	0	0
True 6	1	1	0	1	0	0	16	0
True 7	2	4	3	0	0	0	1	3

1. **Rendimiento General**

El análisis de la matriz de confusión muestra que el modelo tiene un rendimiento desigual entre diferentes clases. La clase 1 tiene el mejor desempeño con 56 verdaderos positivos y solo algunas confusiones menores, mientras que la clase 6 también muestra buenos resultados con 16 verdaderos positivos y mínimas confusiones. Sin embargo, las clases 2, 3 y 4 tienen problemas de clasificación, con múltiples errores de predicción y pocos verdaderos positivos. En general, el modelo parece tener una precisión aceptable en la mayoría de las clases, pero presenta dificultades significativas con algunas, lo que sugiere la necesidad de ajustes o mejoras adicionales.

1. **Curva ROC**

El análisis de los resultados de AUC-ROC muestra un desempeño sólido del modelo en la mayoría de las clases. La clase 5 tiene la puntuación más alta con un AUC de 0.994, indicando una excelente capacidad para distinguir entre esa clase y las demás. La clase 6 también muestra un AUC muy alto de 0.982, reflejando un buen rendimiento en la clasificación de esa categoría. Las clases 0, 1, 3 y 4 también tienen valores altos, superiores a 0.90, mientras que la clase 2 y la clase 7, con AUCs de 0.834 y 0.878 respectivamente, presentan un rendimiento algo inferior. El AUC promedio macro de 0.918 sugiere que, en general, el modelo ofrece un buen equilibrio en la capacidad de discriminación entre todas las clases.

1.2.1. Comparación General de los tres modelos

- **Modelos y AUC**:
  - El análisis de la matriz de confusión muestra que el modelo actual tiene un rendimiento variado entre las clases, con la clase 5 destacándose con un AUC de 0.994 y la clase 2 mostrando un rendimiento inferior con un AUC de 0.834. Este modelo presenta un buen desempeño general, aunque con áreas que requieren mejoras.
  - En comparación con CancerSEEK, que exhibe un rendimiento polarizado con buenos resultados en ciertos tipos de cáncer pero deficiencias en otros, el modelo actual logra un rendimiento alto en la mayoría de las clases, aunque presenta debilidades en algunas categorías específicas.
  - CancerSEEK tiene una polarización notable en su rendimiento entre diferentes tipos de cáncer, con buenos resultados en cánceres colorrectal y de ovario pero bajo rendimiento en tipos como el cáncer de hígado, indicando una posible limitación en la cobertura de datos.
- **Uniformidad y Diversidad**:
  - El modelo actual muestra alta uniformidad con un AUC promedio macro de 0.918, sugiriendo un equilibrio general en el rendimiento entre las clases. Sin embargo, hay una notable diversidad, especialmente con las clases 2 y 7, que presentan puntuaciones más bajas.
  - CancerSEEK muestra una mayor variabilidad en el rendimiento, con polarización que indica una falta de uniformidad, probablemente relacionada con la disponibilidad desigual de datos para los diferentes tipos de cáncer.
  - En contraste, los resultados del análisis de probabilidades predichas muestran que el modelo es más consistente en su capacidad de identificación para clases como **Liver** y **Lung**, mientras que muestra confusión en clases como **Ovary** y **Pancreas**.

- **Top Performers**:
  - El modelo actual identifica de manera sobresaliente las clases 5 y 6, con AUCs superiores a 0.98, indicando un excelente rendimiento en la clasificación de estas categorías. También muestra un buen desempeño en **Breast** y **Colorectum**.
  - CancerSEEK, aunque efectivo en algunos cánceres, tiene problemas significativos en tipos menos representados como el cáncer de hígado, reflejando una capacidad sesgada hacia ciertos tipos de cáncer.
  - En el análisis de probabilidades, **Liver** y **Lung** tienen las probabilidades predichas más altas, lo que sugiere un buen desempeño en estas categorías, mientras que **Pancreas** y **Ovary** presentan dificultades en diferenciación clara.
- **Rendimiento Inferior**:
  - El modelo actual muestra un rendimiento más bajo en las clases 2 y 7, con AUCs de 0.834 y 0.878, sugiriendo áreas de mejora necesarias para estas categorías.
  - CancerSEEK también enfrenta problemas en tipos de cáncer menos representados, como el cáncer de hígado, indicando limitaciones relacionadas con datos insuficientes para algunas clases.
  - El análisis de probabilidades predichas revela que **Esophagus** y **Stomach** tienen distribuciones de probabilidades que se extienden a otras categorías, sugiriendo que el modelo podría estar confuso o tener una menor capacidad discriminativa en estas áreas.
- **Interpretación y Conclusiones**:
  - El modelo actual muestra un desempeño sólido en general con áreas para mejorar en las clases con menor AUC, indicando una capacidad aceptable para diferenciar entre la mayoría de las clases. Ajustar el modelo podría ser beneficioso para mejorar el rendimiento en las categorías con puntuaciones más bajas.
  - CancerSEEK presenta un buen rendimiento en ciertos cánceres pero revela una polarización en su capacidad de clasificación, lo que subraya la necesidad de equilibrar el conjunto de datos o ajustar el modelo para mejorar la uniformidad en el rendimiento.
  - El análisis de probabilidades predichas confirma que el modelo es efectivo en identificar ciertos tipos de cáncer como **Liver** y **Lung**, pero presenta oportunidades para mejorar la precisión en la clasificación de clases con distribuciones más dispersas. Revisar y ajustar el enfoque de predicción podría mejorar la diferenciación entre clases menos claramente separadas.

Despliegue Tecnológico

Plan de Despliegue

El despliegue tecnológico de nuestro modelo de detección de cáncer es una etapa crítica para asegurar que el sistema funcione correctamente en un entorno real. Este plan detallado se divide en varias fases, cada una diseñada para minimizar riesgos y asegurar un funcionamiento eficiente y sin interrupciones.

Fases del Despliegue

El despliegue se llevará a cabo en las siguientes fases:

1. **Entrega Inicial de Verificación**

Objetivos:

- Verificación de la integración del software.
- Comprobación de las comunicaciones y acceso a ficheros.
- Realización de fine tuning necesarios.

Actividades:

- **Verificación de Software y Comunicaciones**: Asegurar que el modelo interactúa correctamente con el sistema existente, verificando que no hay problemas de compatibilidad.
- **Acceso a Ficheros**: Garantizar que el modelo puede acceder a todos los datos necesarios, realizar operaciones de lectura y escritura sin problemas.
- **Fine Tuning**: Ajustar parámetros del modelo y del sistema para optimizar el rendimiento en el entorno de producción.

Ventajas:

- Identificación temprana de problemas potenciales.
- Reducción de riesgos antes de la implementación final.

2. **Entrega Definitiva**

Objetivos:

- Implementar la versión final del modelo.
- Verificación final de cambios y correcciones.

Actividades:

- **Subida del Modelo Final**: Implementar el modelo optimizado y completamente entrenado en el entorno de producción.
- **Verificación de Cambios**: Asegurar que todas las mejoras y correcciones identificadas en la fase inicial han sido implementadas correctamente.

Ventajas:

- Asegurar que el modelo está listo para su uso por los usuarios finales.
- Validación de que el modelo cumple con todos los requisitos de funcionamiento.

Capacitación del Personal

Objetivos:

- Garantizar que el personal que operará el sistema está adecuadamente capacitado.
- Proveer documentación detallada para el uso y mantenimiento del modelo.

Actividades:

- **Formación Específica**: Entrenamiento en el uso del modelo, interpretación de resultados, y manejo de posibles errores o anomalías.
- **Documentación**: Proveer manuales y guías detalladas sobre el funcionamiento del modelo y los procedimientos de mantenimiento.

Ventajas:

- Asegurar un manejo eficiente del sistema.

- Minimizar tiempos de inactividad por errores operativos.

Riesgos y Mitigaciones

Riesgos Identificados:

- **Incompatibilidad de Software**: Posibles conflictos entre el modelo y el software existente.
- **Problemas de Comunicación**: Fallos en la transmisión de datos entre componentes del sistema.
- **Errores de Acceso a Datos**: Problemas al leer o escribir datos necesarios para el modelo.
- **Falta de Capacitación**: Personal no capacitado que podría causar errores operativos.

Estrategias de Mitigación:

- **Pruebas Exhaustivas en la Entrega Inicial**: Realizar pruebas detalladas de integración y comunicación para identificar y resolver problemas antes de la implementación completa.
- **Capacitación y Documentación**: Asegurar que todo el personal esté bien entrenado y tenga acceso a documentación detallada.
- **Monitorización Continua**: Supervisar el rendimiento del modelo en tiempo real para detectar y corregir rápidamente cualquier problema que surja.

Evaluación y Presentación de Resultados

Objetivos:

- Evaluar continuamente el rendimiento del modelo.
- Presentar los resultados y beneficios a todas las partes interesadas.

Actividades:

- **Monitoreo Continuo**: Supervisar el modelo para asegurar su correcto funcionamiento y realizar ajustes según sea necesario.
- **Evaluación Periódica**: Revisar periódicamente el rendimiento del modelo para asegurar que sigue cumpliendo con los objetivos.
- **Presentación de Resultados**: Compartir resultados y beneficios obtenidos con todas las partes interesadas, destacando el impacto positivo del modelo en la detección temprana del cáncer.

Ventajas:

- Asegurar la mejora continua del sistema.
- Mantener a todas las partes interesadas informadas y alineadas con los objetivos del proyecto.

Conclusión

El plan de despliegue está diseñado para asegurar una implementación exitosa y sin interrupciones del modelo de detección de cáncer. A través de una entrega inicial de verificación, una entrega definitiva, capacitación del personal, y una evaluación continua, buscamos minimizar los riesgos y maximizar los beneficios del sistema. Este enfoque garantiza que el modelo no solo se implemente eficazmente, sino que también se mantenga y mejore continuamente, proporcionando un valor significativo en la detección temprana y el tratamiento del cáncer.

Puesta en Valor

Plan de Despliegue Operativo

En esta sección se presentará la estrategia para desplegar los resultados analíticos obtenidos del modelo de predicción de cáncer en los procesos de las entidades interesadas. Este plan incluye tanto la preparación de los modelos entrenados como la oferta de estos modelos a terceros para su implementación en entornos productivos.

Objetivo

El objetivo principal del despliegue es proporcionar a las entidades interesadas, como hospitales, clínicas y laboratorios de investigación, una herramienta eficiente y precisa para la detección temprana del cáncer a partir de biomarcadores en análisis de sangre. Esto permitirá a los profesionales de la salud tomar decisiones más informadas y mejorar los resultados de los pacientes.

Plan de Despliegue

El despliegue se realizará en varias fases para garantizar un proceso fluido y controlado, facilitando la integración con los sistemas de las entidades receptoras.

1. **Fase de Preparación**
    - **Verificación Inicial**: Se llevará a cabo una revisión exhaustiva del código y los modelos para asegurar que están libres de errores y listos para ser entregados.
    - **Documentación Completa**: Preparación de documentación detallada, incluyendo manuales de usuario, guías de integración y especificaciones técnicas.
2. **Fase de Oferta**
    - **Identificación de Entidades Interesadas**: Identificar y contactar a hospitales, clínicas y laboratorios que puedan beneficiarse del uso de los modelos.
    - **Presentación de la Solución**: Realización de presentaciones y demostraciones para mostrar los beneficios y la precisión del modelo, destacando su capacidad para mejorar la detección temprana del cáncer.
3. **Fase de Implementación**
    - **Entrega de Modelos**: Proporcionar los modelos entrenados junto con la documentación necesaria a las entidades receptoras.
    - **Asesoramiento en la Integración**: Ofrecer soporte técnico y asesoramiento durante el proceso de integración de los modelos en los sistemas de las entidades.
4. **Fase de Pruebas**
    - **Pruebas de Funcionamiento**: Asistir a las entidades en la realización de pruebas exhaustivas para asegurar que el modelo funciona correctamente en su entorno productivo.
    - **Pruebas de Usuario**: Facilitar un periodo de prueba donde los usuarios finales puedan interactuar con el modelo y proporcionar retroalimentación.
5. **Fase de Monitoreo y Mantenimiento**
    - **Monitoreo Continuo**: Implementar herramientas de monitoreo para supervisar el desempeño del modelo y la aplicación, asegurando la detección temprana de cualquier problema.
    - **Actualizaciones y Mejoras**: Planificación de un ciclo regular de actualizaciones para incorporar mejoras basadas en la retroalimentación de los usuarios y los avances en la investigación.

Ventajas

- **Detección Temprana y Precisa**: La implementación de este modelo permitirá la detección temprana del cáncer, mejorando las tasas de supervivencia y reduciendo los costos de tratamiento.
- **Optimización de Recursos**: Al automatizar el proceso de análisis y predicción, se optimiza el uso de recursos médicos y se reduce la carga de trabajo para los profesionales de salud.
- **Accesibilidad**: Las entidades receptoras podrán integrar fácilmente los modelos en sus sistemas, mejorando la accesibilidad a diagnósticos tempranos y precisos.

Riesgos y Mitigaciones

- **Riesgo de Inexactitud en las Predicciones**: Aunque el modelo ha sido entrenado y validado exhaustivamente, siempre existe el riesgo de predicciones inexactas. Para mitigar esto, se implementarán mecanismos de doble verificación y se proporcionará una guía clara sobre cómo interpretar los resultados.
- **Riesgo de Seguridad de Datos**: La protección de datos sensibles de los pacientes es crucial. Se implementarán medidas de seguridad robustas, incluyendo cifrado de datos y protocolos de acceso seguro.
- **Resistencia al Cambio**: Algunos profesionales de la salud pueden ser reacios a confiar en sistemas automatizados. Se planificarán sesiones de formación y demostraciones para mostrar los beneficios y la fiabilidad del sistema.

Aplicación Complementaria

Además del modelo principal, se ha desarrollado una aplicación complementaria que permite a los usuarios introducir parámetros de análisis de sangre y seleccionar el modelo de predicción para obtener resultados inmediatos. La aplicación no es el objetivo final del proyecto, pero añade un valor significativo al permitir un acceso más fácil y rápido a las predicciones.

- **Funcionalidades de la Aplicación**:
  - Entrada de Parámetros: Los usuarios pueden introducir datos de análisis de sangre.
  - Selección de Modelos: Posibilidad de seleccionar entre diferentes modelos de predicción.
  - Resultados de Predicción: La aplicación muestra si el paciente tiene cáncer o no, basado en los datos introducidos.

Conclusión

La puesta en valor de este proyecto mediante la oferta de los modelos entrenados a entidades interesadas y la implementación de una aplicación complementaria no solo mejorará la detección temprana del cáncer, sino que también optimizará los recursos médicos y aumentará la accesibilidad a diagnósticos rápidos y precisos. Este enfoque asegura que los avances analíticos se traduzcan en beneficios clínicos tangibles, alineándose con los objetivos establecidos en la fase de Comprensión de Negocio y Evaluación y Presentación de Resultados.

Conclusiones

Resumen de Objetivos Alcanzados

El proyecto ha logrado varios objetivos clave en el desarrollo de modelos de machine learning para la detección del cáncer a partir de biomarcadores en análisis de sangre. Entre los logros más destacados se encuentran:

1. **Desarrollo de Modelos Supervisados**: Se implementaron diversos modelos de aprendizaje supervisado como regresión logística, Random Forest, KNN, AdaBoost, Gradient Boosting y Voting Classifier. Estos modelos se evaluaron exhaustivamente, y se determinó que algunos, como Random Forest y Gradient Boosting, ofrecían los mejores resultados en términos de precisión y capacidad predictiva.
2. **Validación Cruzada**: Se aplicó la validación cruzada para asegurar que los modelos generalicen bien a nuevos datos. Se utilizaron métricas como precisión, recall y F1-score para evaluar el rendimiento de los modelos.
3. **Desarrollo de Aplicación**: Se creó la aplicación 'CancerDetector.py', que permite a los usuarios introducir parámetros de análisis de sangre y seleccionar un modelo para predecir la presencia de cáncer. La aplicación también incluye una pestaña para predecir el tipo de cáncer utilizando un conjunto de modelos combinados mediante Voting Classifier.
4. **Análisis de Modelos No Supervisados**: Aunque se exploraron técnicas de aprendizaje no supervisado como KMeans, DBSCAN y GMM, se concluyó que estos modelos no eran adecuados para el tipo de datos y el objetivo del proyecto, ya que no aportaban mejoras significativas en la predicción del cáncer.

Análisis Crítico

Calidad y Cantidad de Datos

- **Calidad de los Datos**: La calidad de los datos utilizados en este estudio es alta, con biomarcadores seleccionados específicamente para la detección del cáncer. Sin embargo, la muestra es pequeña, lo que limita la capacidad de los modelos para generalizar y puede llevar a problemas de sobreajuste.
- **Generación de Datos Sintéticos**: Aunque se utilizó el modelo CTGAN para generar datos sintéticos y aumentar la muestra, este enfoque puede introducir sesgos y no replicar completamente las complejas relaciones entre los biomarcadores.

Modelos y Resultados

- **Modelos Supervisados**: Los modelos supervisados demostraron ser eficaces para la predicción del cáncer, con Random Forest y Gradient Boosting destacándose como los más precisos. Sin embargo, la variabilidad en las predicciones sugiere que hay espacio para mejorar en la optimización de hiperparámetros y en la recolección de más datos.
- **Modelos No Supervisados**: Los modelos no supervisados no proporcionaron mejoras significativas. Esto se debe a la naturaleza específica de los datos y a la necesidad de predicciones precisas en lugar de agrupaciones o patrones generales.

Conclusiones Generales del Proyecto

Muestra de Datos

Ampliación de la Muestra de Datos: La recolección de más datos reales es esencial para mejorar la robustez y la confiabilidad de los modelos predictivos. Actualmente, los datos disponibles son de alta calidad pero insuficientes en cantidad. Ampliar la muestra permitirá que los modelos generalicen mejor y reduzcan el riesgo de sobreajuste. La integración de datos de diferentes cohortes, etnias y condiciones clínicas proporcionará una visión más completa y generalizable. Además, la aplicación de técnicas de augmentación de datos, como la generación de variaciones de los datos existentes, puede simular diferentes condiciones y aumentar la diversidad del conjunto de datos sin necesidad de recolectar nuevos datos.

Datos Sintéticos como Complemento: Aunque los datos reales son ideales, los datos sintéticos pueden complementar significativamente la muestra existente. La utilización de modelos generativos como CTGAN para generar datos sintéticos ha mostrado potencial para replicar distribuciones de datos reales. Sin embargo, es crucial validar estos datos para asegurar que mantengan la integridad y las relaciones entre variables, evitando sesgos que puedan afectar negativamente el rendimiento del modelo.

Optimización de Modelos

Ajuste de Hiperparámetros: La optimización continua de los hiperparámetros de los modelos es vital para mejorar su rendimiento. Técnicas como Grid Search y Random Search pueden ayudar a identificar las configuraciones óptimas. Además, explorar nuevas arquitecturas de modelos, incluyendo redes neuronales más profundas y avanzadas, podría ofrecer mejoras significativas en la precisión de las predicciones.

Modelos de Ensamble: Agregar más modelos al Voting Classifier puede aumentar la precisión y la robustez de las predicciones. Los modelos de ensamble combinan las fortalezas de múltiples modelos individuales, reduciendo el riesgo de sobreajuste y mejorando la capacidad de generalización. Continuar explorando y ajustando estos modelos puede ofrecer beneficios sustanciales.

Aplicación Práctica

Herramienta 'CancerDetector.py': La aplicación desarrollada, 'CancerDetector.py', es una herramienta práctica y útil para poner a prueba los modelos entrenados. Permite a los usuarios experimentar con diferentes modelos y ver cómo los distintos biomarcadores afectan las predicciones. Esta aplicación no solo facilita la validación de los modelos en un entorno controlado, sino que también sirve como un demostrador tecnológico para posibles colaboraciones y despliegues futuros.

Interactividad y Flexibilidad: La posibilidad de seleccionar diferentes modelos y ajustar parámetros permite a los usuarios entender mejor cómo funcionan los modelos y qué factores son más influyentes en las predicciones. Esta interactividad es clave para educar a los usuarios sobre la detección del cáncer y la importancia de los biomarcadores.

Enfoque en la Calidad

Mantenimiento de la Calidad de los Datos: Dado que los datos son de alta calidad pero en pequeña cantidad, es fundamental mantener este enfoque en la calidad en futuras recolecciones. La precisión y la integridad de los datos son cruciales para entrenar modelos fiables. A medida que se recolecten más datos, es importante implementar rigurosos procesos de limpieza y validación para asegurar que la calidad se mantenga alta.

Diversidad de los Datos: La inclusión de datos de diversas poblaciones ayudará a crear modelos más equitativos y aplicables a diferentes grupos demográficos. Esto no solo mejorará la precisión del modelo, sino que también garantizará que las predicciones sean relevantes y útiles para una amplia gama de pacientes.

Relevancia de los Biomarcadores

Validación Continua de Biomarcadores: Los biomarcadores seleccionados han demostrado ser efectivos en la predicción del cáncer. Sin embargo, es crucial seguir investigando y validando estos biomarcadores en estudios futuros. La constante validación ayudará a confirmar su relevancia y descubrir nuevos marcadores que puedan mejorar aún más las predicciones.

Investigación y Descubrimiento de Nuevos Marcadores: Continuar la investigación para identificar nuevos biomarcadores que puedan proporcionar información adicional sobre la presencia de cáncer. La combinación de biomarcadores tradicionales con nuevos descubrimientos puede llevar a modelos más precisos y robustos.

Dificultades en el Desarrollo del Proyecto

Durante el desarrollo del proyecto para la detección temprana de cáncer a partir de pruebas de sangre multianalíticas, se han identificado diversas dificultades significativas que pueden afectar el éxito y la implementación del proyecto. A continuación, se detallan estos desafíos:

Falta de Datos Reales

Escasez de Datos Auténticos y Diversos: La falta de datos representativos y en cantidad suficiente dificulta la evaluación del desempeño de los modelos en escenarios reales. Esto afecta la precisión de las predicciones y limita la capacidad de los modelos para generalizar a nuevos datos. Además, la ausencia de datos específicos de diferentes poblaciones puede impedir que el modelo sea equitativo y aplicable a diversas subpoblaciones.

Diversificación de la Muestra: Es crucial ampliar no solo la cantidad de datos, sino también su diversidad, incluyendo datos de diferentes cohortes, etnias y condiciones clínicas. Esto proporcionará una visión más completa y generalizable de los patrones relevantes para la detección del cáncer.

Data Augmentation: Considerar técnicas específicas de augmentación de datos para biomarcadores, como la generación de variaciones en los datos existentes, para simular diferentes condiciones sin necesidad de recolectar más datos.

Integración de Datos Genómicos y Clínicos: Incorporar datos genómicos, clínicos y de imágenes médicas para obtener una visión más holística y mejorar significativamente la calidad de los modelos predictivos.

Dependencia del Entorno de Desarrollo

Portabilidad y Escalabilidad: La dependencia del entorno de desarrollo, como una máquina virtual (VM), crea problemas de portabilidad y escalabilidad. Los modelos entrenados deben cargarse en el mismo entorno en el que fueron desarrollados, lo que puede dificultar su implementación en diferentes plataformas o contextos operativos. Esto limita la flexibilidad en la gestión y mantenimiento de los modelos.

Predicciones Incorrectas Iniciales

Preprocesamiento de Datos: Las predicciones incorrectas iniciales, como la detección de cáncer en todos los casos, se deben a la falta de preprocesamiento adecuado de los datos de entrada. Sin un preprocesamiento consistente, las discrepancias entre los datos de entrada y los datos de entrenamiento pueden llevar a resultados incorrectos. Esto subraya la necesidad de un pipeline de preprocesamiento robusto que garantice la compatibilidad de los datos con el modelo.

Limitaciones de Recursos Computacionales

Recursos Computacionales: El entrenamiento de modelos complejos requiere una cantidad significativa de recursos computacionales, incluyendo tiempo de procesamiento, memoria y almacenamiento. Las limitaciones en estos recursos pueden ralentizar el desarrollo e implementación de los modelos, aumentando los costos y afectando la eficiencia general del proyecto. Esto es especialmente desafiante si no se dispone de infraestructura avanzada, como clústeres de computación o servicios en la nube.

Complejidad en la Interpretación de Resultados

Modelos Complejos y Caja Negra: Los modelos complejos, como las redes neuronales profundas, a menudo se consideran "caja negra", lo que dificulta su interpretación y justificación. En aplicaciones médicas, la transparencia es crucial para que los resultados sean comprensibles y justificables para los profesionales de la salud. La falta de interpretabilidad puede generar desconfianza en los resultados y limitar la adopción de los modelos.

Gestión de la Calidad de los Datos

Calidad de los Datos: La calidad de los datos de entrada es esencial para el desarrollo de modelos fiables. Datos incompletos, ruidosos o sesgados pueden llevar a resultados incorrectos y a modelos poco fiables. La limpieza de datos, la imputación de valores nulos, la normalización y estandarización, y la eliminación de valores atípicos son procesos críticos que requieren atención meticulosa para asegurar la precisión del modelo.

Evolución de los Datos y el Modelo

Cambio de Datos y Condiciones: Los datos y las condiciones pueden cambiar con el tiempo debido a diversos factores, como nuevas tecnologías de diagnóstico o tratamientos, y cambios en la población estudiada. Estos cambios pueden afectar la precisión y relevancia de los modelos entrenados. Es necesario implementar estrategias de mantenimiento y actualización continua para garantizar que los modelos sigan siendo precisos y útiles.

Consideraciones Éticas y de Privacidad

Normativas de Privacidad y Ética: El manejo de datos sensibles, especialmente en el contexto de la salud, requiere cumplir con normativas de privacidad y ética, como el GDPR en Europa y la HIPAA en Estados Unidos. Estas regulaciones pueden complicar el acceso a ciertos conjuntos de datos y limitar el análisis completo sin comprometer la privacidad de los pacientes. Es fundamental abordar estas consideraciones éticas para garantizar el uso responsable de los datos.

Integración con Sistemas Existentes

Compatibilidad e Integración: Integrar modelos en sistemas y flujos de trabajo existentes puede ser desafiante. Los problemas pueden incluir incompatibilidades de formato de datos y limitaciones en el rendimiento de los sistemas actuales. La integración exitosa requiere una planificación cuidadosa y posiblemente la adaptación de sistemas para garantizar que los modelos se implementen de manera eficiente.

Evaluación y Validación del Modelo

Generalización y Sobreajuste: Garantizar que el modelo generalice bien a nuevos datos y no esté sobreajustado a los datos de entrenamiento es un desafío constante. Técnicas robustas de validación, como la validación cruzada y el uso de conjuntos de datos de prueba independientes, son necesarias para garantizar la fiabilidad del modelo. También es crucial realizar un monitoreo continuo del rendimiento para identificar y corregir cualquier disminución en su precisión o efectividad.

Desafíos en la Recolección de Datos

Heterogeneidad y Calidad: La recolección de datos puede enfrentar desafíos relacionados con la heterogeneidad de las fuentes de datos, la calidad variable de los datos recolectados y las dificultades en el acceso a datos específicos. La coordinación con diferentes instituciones para obtener datos de calidad y asegurar la consistencia en la recolección es fundamental para obtener un conjunto de datos robusto y representativo.

Desafíos en la Explicación y Comunicación de Resultados

Comunicación Efectiva: Comunicar los resultados del modelo a las partes interesadas, incluyendo profesionales médicos y pacientes, puede ser complicado. La necesidad de traducir resultados técnicos en información comprensible para audiencias no técnicas es un desafío importante. Además, garantizar que las recomendaciones del modelo se alineen con las prácticas clínicas y los protocolos existentes requiere una colaboración efectiva con expertos en el dominio médico.

Desafíos en la Adaptación a Nuevas Variantes del Cáncer

Nuevas Variantes y Subtipos: La detección temprana de cáncer puede complicarse con la aparición de nuevas variantes o subtipos de cáncer. Los modelos pueden necesitar ajustes y reentrenamiento para abordar eficazmente estas nuevas variantes, lo que requiere un enfoque dinámico y flexible en el desarrollo del modelo.

Desafíos en la Implementación en Entornos Clínicos

Interoperabilidad y Aceptación: La implementación de modelos en entornos clínicos puede enfrentar desafíos relacionados con la interoperabilidad con sistemas de información médica existentes, la aceptación por parte de los profesionales de la salud y la integración en los flujos de trabajo clínicos. Es necesario un enfoque colaborativo para garantizar que el modelo se adapte bien a las prácticas clínicas y aporte valor a los procesos de diagnóstico y tratamiento.

Caminos Abiertos y Próximos Pasos

Calidad y Cantidad de Datos

Ampliación:

Diversificación de la Muestra: Es fundamental ampliar no solo la cantidad de datos sino también su diversidad, incluyendo datos de diferentes cohortes, etnias y condiciones clínicas. Esto permitirá obtener una visión más completa y generalizable de los patrones relevantes para la detección del cáncer.

Data Augmentation: Considerar técnicas específicas de augmentación de datos para biomarcadores, como la generación de variaciones en los datos existentes. Esto ayudará a simular diferentes condiciones sin necesidad de recolectar más datos, aumentando la robustez del modelo.

Nuevos Puntos:

Integración de Datos Genómicos y Clínicos: Incorporar datos genómicos, clínicos y de imágenes médicas para obtener una visión más holística. Esto mejorará significativamente la calidad de los modelos predictivos al proporcionar una imagen completa de los factores involucrados en la detección del cáncer.

Actualización Continua de Datos: Implementar un sistema para la recolección continua de datos, permitiendo la actualización y refinamiento de los modelos en tiempo real, adaptándose así a nuevos hallazgos y cambios en las tendencias clínicas.

Generación de Datos Sintéticos

Ampliación:

Evaluación de Diferentes Técnicas de Generación: Comparar CTGAN con otras técnicas de generación de datos sintéticos, como los Modelos Generativos Adversariales (GANs) o los Modelos Variacionales, para determinar cuál ofrece una mejor replicación de los datos originales.

Nuevos Puntos:

Validación de Datos Sintéticos: Implementar métodos de validación cruzada para asegurar que los datos sintéticos generados mantengan la integridad y la distribución de las relaciones entre variables, garantizando su utilidad en el entrenamiento de modelos.

Ensayo de Aplicación en Modelos: Utilizar los datos sintéticos en diferentes etapas del proceso de modelado (pre-entrenamiento, fine-tuning) para evaluar su impacto en el rendimiento del modelo.

Procesamiento de Datos

Ampliación:

Imputación Avanzada: Explorar técnicas avanzadas como la imputación basada en redes neuronales, métodos bayesianos, o imputación múltiple para mejorar el manejo de datos faltantes, aumentando la precisión y la robustez de los modelos.

Nuevos Puntos:

Transformaciones de Datos: Implementar transformaciones de datos como escalado, normalización y técnicas de reducción de dimensionalidad (p.ej., PCA) para optimizar la calidad de los datos antes de su procesamiento, mejorando así el rendimiento del modelo.

Replicación de Modelos de Referencia

Ampliación:

Documentación Detallada: Asegurarse de que toda la documentación del modelo de referencia esté completa, incluyendo los parámetros exactos y el entorno de ejecución. Esto facilitará la replicación y la validación de los resultados.

Nuevos Puntos:

Colaboración con los Creadores Originales: Establecer colaboraciones con los autores de los modelos originales para obtener insights adicionales y asistencia en la replicación, asegurando la fidelidad y el rendimiento del modelo replicado.

Tendencia al Sobreajuste

Ampliación:

Uso de Validación Cruzada Estratificada: Implementar técnicas de validación cruzada estratificada para evaluar la robustez del modelo en diferentes subconjuntos de datos, reduciendo así el riesgo de sobreajuste.

Modelos de Ensamble: Continuar explorando modelos de ensamble, como Random Forests y Gradient Boosting, para combinar fortalezas de diversos modelos y mitigar el riesgo de sobreajuste.

Nuevos Puntos:

Optimización de Hiperparámetros: Utilizar técnicas de optimización automatizada como Grid Search y Random Search para ajustar hiperparámetros de manera más efectiva, mejorando el rendimiento del modelo.

Regularización Avanzada: Implementar técnicas avanzadas de regularización, como Dropout en redes neuronales, para combatir el sobreajuste y aumentar la generalización del modelo.

Ampliación y Refinamiento de Modelos

Ampliación:

Exploración de Modelos Alternativos: Evaluar nuevos modelos emergentes como Transformers y otros enfoques de aprendizaje profundo que podrían ofrecer mejoras en la predicción, proporcionando nuevas perspectivas y técnicas avanzadas.

Nuevos Puntos:

Desarrollo de Modelos Personalizados: Crear modelos personalizados adaptados específicamente a las características de los datos y el problema en cuestión, en lugar de depender únicamente de modelos preexistentes, para optimizar la precisión y la relevancia del modelo.

Evaluación de Modelos No Supervisados

Ampliación:

Análisis de Clustering Avanzado: Realizar un análisis más profundo utilizando técnicas de clustering avanzadas (p.ej., clustering jerárquico, t-SNE) para identificar patrones ocultos en los datos que puedan mejorar la comprensión y predicción del cáncer.

Nuevos Puntos:

Integración con Modelos Supervisados: Combinar resultados de modelos no supervisados con modelos supervisados para mejorar la precisión y la interpretación de las predicciones, aprovechando las fortalezas de ambos enfoques.

Variabilidad en las Predicciones

Ampliación:

Análisis de Sensibilidad: Realizar un análisis de sensibilidad para entender cómo diferentes variables afectan las predicciones y mitigar la variabilidad en las probabilidades asignadas, mejorando la robustez del modelo.

Nuevos Puntos:

Implementación de Técnicas de Ensayo y Error: Introducir técnicas de ensayo y error para evaluar cómo diferentes enfoques y técnicas afectan la variabilidad en las predicciones, refinando así el proceso predictivo.

Mejoras Adicionales

Ampliación:

Optimización de Modelos Existentes: Continuar con el ajuste y optimización de modelos existentes, incorporando nuevas técnicas y herramientas que puedan surgir en el campo del machine learning.

Nuevos Puntos:

Análisis de Resultados en Diferentes Contextos: Evaluar cómo los modelos se comportan en diferentes contextos clínicos y demográficos para asegurar que sean robustos y aplicables a diversas poblaciones.

Conclusión Final de Caminos Abiertos

Para avanzar en el desarrollo de modelos predictivos en cáncer, se deben abordar múltiples frentes simultáneamente: acceso a datos diversificados y actualizados, implementación de técnicas avanzadas de procesamiento y generación de datos, y mejora continua de modelos mediante la exploración de enfoques nuevos y la reducción del sobreajuste. La combinación de estos esfuerzos contribuirá a la creación de modelos más robustos, confiables y aplicables a un rango más amplio de situaciones clínicas.

Próximos Pasos

1. **Recolección de Más Datos**: Buscar colaboraciones con instituciones médicas y de investigación para recolectar más datos de pacientes.
2. **Mejora de la Aplicación**: Integrar más funcionalidades en la aplicación, como la visualización de tendencias y patrones en los biomarcadores.
3. **Estudio de Nuevas Técnicas**: Explorar técnicas de machine learning más avanzadas y su potencial para mejorar la predicción del cáncer.
4. **Validación en el Mundo Real**: Realizar pruebas piloto de los modelos en entornos clínicos para validar su efectividad en la práctica.

En resumen, aunque el proyecto ha alcanzado sus objetivos principales y ha desarrollado herramientas útiles para la detección del cáncer, hay áreas claras para la mejora y la expansión futura. Con más datos y optimizaciones, los modelos pueden volverse aún más precisos y útiles en entornos clínicos.

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. Early Cancer Detection from Multianalyte Blood Test Results. *PMC*. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6548890/**

10. Salazar-Jordan, Revista Nova. (2009). Cuantificación de ADN libre en plasma sanguíneo de voluntarios sanos en una población bogotana. Recuperado de **https://revistas.unicolmayor.edu.co/index.php/nova/article/view/139/279**

11. Alfaro et al. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature Methods*. **https://www.nature.com/articles/nmeth.3138**

12. Crick, F. H. C. (1958). On protein synthesis. *Symposium of the Society for Experimental Biology*.

13. Mi Yang et al. (2020). Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics. *Cell Systems, 11*(4), 352-360. **https://www.cell.com/cell-systems/fulltext/S2405-4712(20)30242-8**

14. Saez-Rodriguez et al. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews Genetics, 17*(8), 470-486. **https://www.nature.com/articles/nrg.2016.69**

15. Wu et al. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *PubMed*. **https://pubmed.ncbi.nlm.nih.gov/31911677/**