Máster en Business Analytics

Tema:

Cáncer de mama y Business Intelligence: hacia una detección precoz y eficiente

Trabajo final de Máster

Autores:

Nérica Giménez Martínez

Nuria Hernández Puig

Sebastián Ariel Jaimovich

Tutores:

Victor Manuel Yeste Moreno

Vicente Castillo Fauli

Junio 2025

Índice

Resumen

En España, acudir al médico forma parte de nuestra cotidianidad, pero en muchas ocasiones no se es plenamente conscientes del coste real que implica la atención sanitaria que se recibe. Los gastos hospitalarios no se limitan únicamente a la consulta médica, sino que engloban todos los recursos que un hospital o centro médico emplea para prestar servicios a la ciudadanía: pruebas diagnósticas, tratamientos, hospitalizaciones, equipamiento, personal sanitario, entre otros gastos.

Según se indica en el informe anual de "Estadística de gasto sanitario" el gasto ascendió en 2023 a 97.661 millones de euros, lo que si comparamos con el PIB español supone aproximadamente el 6,5%. Para ponerlo en perspectiva, sectores clave como el agroalimentario suponen en torno al 8,9% del PIB, por lo que no podemos decir que los costes sanitarios no tienen relevancia alguna. Si consideramos los habitantes españoles se podría decir que el gasto per cápita fue de 2.021 euros.

A pesar de estos datos, la ciudadanía no siempre es consciente de la magnitud económica que representa una sola visita médica o una intervención sanitaria y es llamativo como de forma generalizada se piensa que es una simple vista, por lo que menos aún se es consciente con enfermedades de alto impacto como el cáncer.

Ahora bien, cierto es que existen diferentes tipologías de gastos sanitarios en el caso actual de este ensayo nos centraremos en los gastos médicos que engloban la detección del cáncer de mama.

Actualmente es muy complicado no haber oído hablar del cáncer o bien conocer a alguien que le hizo frente, lamentablemente sigue siendo uno de los responsables de los fallecimientos en nuestro país. Por ello el cáncer es uno de los grupos de enfermedades de mayor importancia en salud pública y en esta ocasión se pondrá el foco en el cáncer de mama.

Cada año se diagnostican millones de casos nuevos de cáncer de mama, y aunque continuamente se invierte en investigación y en tratamientos la detección temprana es un factor clave para el pronóstico y el desarrollo de la enfermedad.

Cierto es que el avance y las inversiones que se han realizado han mejorado la tasa de supervivencia, aun así, el cáncer de mama es el tumor canceroso más frecuente que sufren las mujeres a nivel mundial y una de las principales causas de mortalidad asociada al cáncer.

El objetivo principal de este Trabajo de Fin de Máster (TFM) es doble. Por un lado, se busca modelar un sistema para la detección temprana del cáncer de mama, con el fin de reducir errores diagnósticos. Por otro, se plantea el desarrollo de un modelo de regresión que permita estimar y reducir el gasto hospitalario asociado a esta enfermedad. Esta aproximación no solo puede mejorar la eficiencia del sistema sanitario, sino también facilitar a gobiernos y organizaciones la planificación de recursos, anticipando tendencias y optimizando las estrategias de prevención y tratamiento.

Palabras claves:

"Cáncer de mama", "modelo predictivo", "modelo de clasificación", "Salud", "machine learning", "gasto hospitalario", "hospitalizaciones".

Abstract

In Spain, visiting the doctor is a part of daily life, but in many cases, people are not fully aware of the actual cost involved in the healthcare they receive. Hospital expenses are not limited solely to medical consultations; they encompass all the resources that a hospital or medical center utilizes to provide services to the public: diagnostic tests, treatments, hospital stays, equipment, healthcare personnel, among other costs.

According to the annual report "Health Expenditure Statistics," healthcare spending in 2023 amounted to €97.661 billion, which represents approximately 6.5% of Spain's GDP. To put this into perspective, key sectors such as the agri-food industry contribute around 8.9% to the GDP, so it is clear that healthcare costs are of significant relevance. Considering the population of Spain, the estimated per capita healthcare expenditure was €2,021.

Despite these figures, citizens are not always aware of the economic magnitude of a single medical visit or intervention. It is striking how it is commonly perceived as "just a check-up," making the awareness even lower when it comes to high-impact illnesses like cancer.

It is important to note that there are different types of healthcare expenses. In this particular study, the focus will be on the medical costs associated with breast cancer detection.

Today, it is almost impossible not to have heard of cancer or not to know someone who has faced it. Unfortunately, it continues to be one of the leading causes of death in our country. For this reason, cancer is among the most significant groups of diseases in public health, and in this study, special attention will be given to breast cancer.

Each year, millions of new breast cancer cases are diagnosed. While continuous investments are made in research and treatment, early detection remains a key factor in the prognosis and progression of the disease.

It is true that medical advances and investments have improved survival rates. Nonetheless, breast cancer remains the most common malignant tumor affecting women worldwide and is one of the leading causes of cancer-related mortality.

The main objective of this Master's Thesis (TFM) is twofold. On one hand, it aims to develop a system for the early detection of breast cancer, with the goal of reducing diagnostic errors. On the other hand, it seeks to build a regression model to estimate and reduce hospital expenses associated with this disease. This approach could not only improve the efficiency of the healthcare system but also support governments and organizations in resource planning by anticipating trends and optimizing prevention and treatment strategies.

Keywords

“Breast cancer”, "Predictive model”, "Classification model”, "Health”, "Hospitalizations”, “Hospital admissions”, "Machine learning”, “Hospital expenditure”

1. Introducción

1.1 La salud en el mundo y en España

La salud es un tema que nos aborda a todos y es primordial para el ser humano. Disponer de buena salud es un desafío cada vez más complejo en el mundo que estamos viviendo. La población mundial sigue en aumento (un crecimiento más leve que años anteriores, pero todavía sigue aumentando), el ser humano se ha vuelto más sedentario en esta era informática utilizando las tecnologías de hoy en día. Las mismas generan una sobrecarga de información (por ejemplo, las redes sociales) y nuestros hábitos cada vez utilizan el menor esfuerzo físico o desplazamiento posible. Estos nuevos comportamientos generan el aumento de ciertos problemas de salud, como ser, la salud mental.

Las economías mundiales no son las mismas que antes, sumado a que el COVID alteró los sistemas de salud logrando que la calidad de estos no sea tan buena como en años anteriores por la alta demanda de pacientes durante y el post pandemia.

Producto de esta saturación a la atención médica, se reemplazaron varias consultas presenciales por consultas virtuales a través de un chat, llamada, o una video llamada. Todavía hay muchos procesos y áreas a mejorar, es una gran oportunidad para optimizar y hacer más eficiente los sistemas de salud a nivel de costes y tiempos.

En base a la información (actualizada hasta el 2023), para entender la situación actual en España, se contabilizaron 48 millones de habitantes, 329.251 nacimientos y 464.417 defunciones. (Ministerio de Sanidad, 2024).

El 20% de la población son mayores de 65 años, la esperanza de vida es de 83 años y 8 de cada 10 personas creen que su estado de salud es bueno. (Ministerio de Sanidad, 2024)

Las muertes en mujeres fueron mayoritariamente por enfermedades cerebrovascular, cáncer de mama y colon. En hombres por enfermedades de isquémica del corazón, cáncer de pulmón y colon.
Durante los últimos años se han desarrollado problemas crónicos de salud que están afectando a la población, en el siguiente gráfico se destacan cuáles son los que están afectando principalmente:

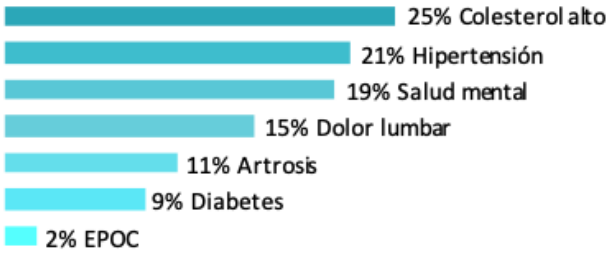


Tabla 1 Principales problemas crónicos de salud, España 2023

Fuente: Ministerio de Sanidad

En cuanto a los estilos de vida, la obesidad afecta al 16% de la población adulta y al 10% en edad infantil. Hay un 68% de gente que consume fruta y un 47% verduras y hortalizas. La actividad física es fundamental en la salud y se sabe que un 36% de las personas no realizan ninguna actividad física.
Por otro lado, los consumos de sustancias tóxicas están afectando a un número importante, el 20% consume tabaco y el 35% consume alcohol a diario. En los últimos 12 meses el 11% de adultos han consumido cannabis (9% mujeres y un 16% hombres) y el 2% cocaína (1% mujeres y un 4% hombres). (Ministerio de Sanidad, 2024)

1.2 El sistema de salud español

Dentro de la población se observa que hay un total de 172.000 profesionales de la Medicina de los cuales 44.000 de ellos se dedican a la atención primaria, 96.000 a la atención hospitalaria, 4.000 a servicios de urgencia y emergencia, como ser la atención telefónica de los servicios 112 y 061. El resto de los 33.000 profesionales son especialistas en formación. (Ministerio de Sanidad, 2024)

En el área de la enfermería hay un total de 227.000 profesionales, contabilizando los ya mencionados 44.000 profesionales de atención primaria, 175.000 en atención hospitalaria, los 4.000 mencionados en servicios de urgencia y emergencia y otros 4.000 profesionales especialistas en formación. (Ministerio de Sanidad, 2024)

Existes otros 382.000 profesionales dedicados a otras ramas de la salud de los cuales 38.000 se dedican a la atención primaria, 324.000 a la atención hospitalaria, 17.000 a los servicios de urgencia y emergencias y otros 3.000 especialistas en formación. (Ministerio de Sanidad, 2024)

La atención sanitaria está compuesta por los centros de atención primaria. Estos son establecimientos públicos conocidos como centros de salud / ambulatorios que cumplen como función brindarle al ciudadano la atención médica básica.

En estos centros hay diversos profesionales, entre ellos los médicos de familia, pediatras, enfermeros/as, matronas, trabajadores sociales, fisioterapeutas, psicólogos y odontología básica.

En total hay 3.000 centros de salud, 10.000 consultorios, 2.000 puntos de atención de urgencia extrahospitalaria, 242 millones consultas médicas, 143 millones consultas de enfermería, 87 millones tele consultas, 13 millones visitas a domicilio y 34 millones de urgencias. (Ministerio de Sanidad, 2024)

Otra rama de la atención sanitaria son las Urgencias y Emergencias 112/061. El número telefónico 112 se estableció en toda Europa para atender emergencias de todo tipo, como ser: sanitarias, incendios, accidentes, delitos, etc. Funciona las 24 horas y es gratuito. El número 061 atiende emergencias sanitarias graves en algunas comunidades autónomas. Para estos servicios están destinados 3.000 ambulancias, hay 8 millones de demandas asistenciales, 600 ambulancias movilizadas por cada 1.000 demandas asistenciales. (Ministerio de Sanidad, 2024)

En la atención sanitaria española también están los hospitales, son centros sanitarios que ofrecen atención médica más compleja y especializada. Comprenden diferentes áreas y servicios de urgencias, especialidades médicas (cardiología, neurología, oncología, etc.), cirugía (operaciones programadas o de urgencia), hospitalización (para tratamientos o recuperaciones) y pruebas de diagnósticos (análisis, radiografías, resonancias, etc.).

Hay aproximadamente 770 centros entre privados y públicos (los primeros superan un poco más en cantidad), estos aproximadamente 159.000 camas disponibles y trabajan más de 57.000 profesionales sanitarios y 157.000 trabajadores proveedores de servicios.

En el año 2023, entre los centros hospitalarios con mejor reputación corporativa del país se destacaron el Hospital Universitario La Paz y la Clínica Universidad de Navarra, en el ámbito público y privado, respectivamente.

Otro factor importante en el sistema de salud es la opinión ciudadana sobre el funcionamiento del mismo. Según el ministerio de sanidad, 6 de cada 10 personas consideran que el sistema sanitario funciona bien, menos de 3 de cada 10 personas opinan que necesita cambios y más de 1 de cada 10 personas piensa que hay que rehacerlo.

La opinión se puede analizar mejor con la valoración de los niveles asistenciales del Sistema Nacional de Salud.

Tabla 2 Valoración de los niveles asistenciales del Sistema Nacional de Salud, España 2023

Fuente: Ministerio de Sanidad

1.2 Gastos sanitarios en España

El gasto sanitario público en el año 2023 en España fue de 97.661 millones de euros, equivalente al 6,5% del producto bruto Interno (PIB). También se notó un ascenso del gasto per cápita a 2.021 euros por habitante (comparado con años anteriores). (Rodríguez Blas, 2025)

A continuación, se detalla la comparativa desde el 2019 hasta el año 2023:

	2019	2020	2021	2022	2023
Millones de euros	74.983	83.630	87.932	91.974	97.661
Porcentaje sobre PIB	6,0%	7,4%	7,1%	6,7%	6,5%
Euros por habitante	1.592	1.766	1.857	1.925	2.021

Tabla 3 Gasto sanitario público consolidado

Fuente: Ministerio de Sanidad

En cuanto al gasto sanitario privado existe información del 2022. Para ese año se gastaron 34.676 millones de euros los cuales representan el 2,6% del PIB y se gastaron 726 euros por habitante. (Rodríguez Blas, 2025)

Los gastos sanitarios se pueden detallar según su clasificación para entender de dónde provienen cada uno de ellos. Los servicios hospitalarios y especializados son los que predominan en gastos con un 61% total consolidado en el año 2023, seguido de Farmacia y de los servicios primarios de Salud con un 14,3%.

En la siguiente tabla se observan los gastos durante los últimos 5 años hasta el 2023, con una tendencia en aumento:

	2019	2020	2021	2022	2023
Servicios hospitalarios y especializados	46.710	52.035	54.268	55.892	59.545
Servicios primarios de salud	10.930	11.881	12.671	13.109	13.962
Servicios de salud pública	823	1.788	2.559	3.023	1.704
Servicios colectivos de salud	1.972	2.065	2.197	2.429	2.614
Farmacia	11.855	12.184	12.855	13.553	13.984
Traslado, prótesis y aparatos terapéuticos	1.320	1.301	1.416	1.542	1.701
Gasto de capital	1.373	2.376	1.965	2.425	4.152
Total consolidado	74.983	83.630	87.932	91.974	97.661

Tabla 4 Gasto sanitario público consolidado según clasificación funcional.

Fuente: Ministerio de Sanidad

A nivel de clasificación sectorial, las Comunidades Autónomas fueron responsables del 93,2% del total del gasto sanitario, las Mutualidades de funcionarios 2,5%, la Seguridad Social 2%, la Administración Central 1,5% y las Corporaciones Locales 0,9%. (Rodríguez Blas, 2025)

Si analizamos el gasto sanitario público del 2023 en cada comunidad autónoma, tuvo un total de 91.004 millones de euros, lo que representa el 6,1% del PIB. El gasto per cápita medio fue de 1.890 euros por habitante. El 46,2% corresponde a las comunidades autónomas de Cataluña, Andalucía y la Comunidad de Madrid. La Rioja, Cantabria y la Comunidad Foral de Navarra son las comunidades con el gasto más bajo, en valores absolutos.

En la siguiente tabla se puede apreciar cuáles fueros los gastos sanitarios en cada comunidad Española:

	Millones de euros	Porcentaje sobre PIB	Euros por habitante
Andalucía	14.178	7,1%	1.648
Aragón	2.696	5,8%	2.004
Asturias (Principado de)	2.196	7,8%	2.182
Balears (Illes)	2.248	5,3%	1.839
Canarias	4.365	8,1%	1.962
Cantabria	1.214	7,2%	2.063
Castilla y León	4.937	7,0%	2.070
Castilla-La Mancha	3.946	7,3%	1.886
Cataluña	15.963	5,7%	2.006
Comunitat Valenciana	9.524	6,8%	1.810
Extremadura	2.215	8,9%	2.101
Galicia	5.365	6,9%	1.988
Madrid (Comunidad de)	11.919	4,1%	1.719
Murcia (Región de)	3.325	8,2%	2.134
Navarra (Comunidad Foral de)	1.411	5,6%	2.092
País Vasco	4.903	5,6%	2.208
Rioja (La)	599	5,6%	1.853
Total Comunidades Autónomas	91.004	6,1%	1.890

Tabla 5 Gasto sanitario público consolidado según comunidad autónoma

Fuente: Ministerio de Sanidad

Comparando con el año anterior, el gasto total (incluyendo la inversión, gastos de capital), el sector Comunidades Autónomas sufrió un aumento del 8,1%. Las comunidades que más incrementaron su gasto fueron la Región de Murcia con un 10,1%, Cataluña con un 10% y la Comunidad Valenciana con un 9,9%. (Rodríguez Blas, 2025)

En las comunidades autónomas, uno de los gastos más significativos es la remuneración del personal. En el año 2023 alcanzó un costo global de 41.095 millones de euros, lo que representa el 45,2% del gasto consolidado del sector, y un incremento de un 6,5% respecto al año anterior.

1.4 Costes generales relacionados con el cáncer

1.4.1 Datos relevantes

A nivel global, el cáncer se reconoce como un problema sociosanitario de gran magnitud y se posiciona como la segunda causa de muerte a nivel mundial. Se estima que para que el año 2030 habrá un aumento de nuevos casos de cáncer superior al 30%.

La Organización Mundial de la Salud (OMS) estima que entre el 30% y el 50% de los casos de cáncer podrían evitarse mediante la adopción de estilos de vida saludables y la implementación de medidas de salud pública efectivas.
Consideramos estos datos altamente relevantes para confirmar que hay una gran oportunidad de mejora en hacer más eficientes los procesos que abordan éste área.

Otro dato no menor es que España cuenta con más médicos pero menos enfermeros que la mayoría de los países de la Unión Europea. En 2022, España tenía 773 médicos por cada 1000 nuevos casos de cáncer versus 679 de media en la Unión Europea.

En el caso de los enfermeros (para atenciones relacionadas con el cáncer), España cuenta con 1106 enfermeros por cada 1000 nuevos casos y la media de la Unión Europea es de 1376 por cada 1000 casos.

En el siguiente gráfico se puede apreciar cómo están distribuidos los recursos en cada país:

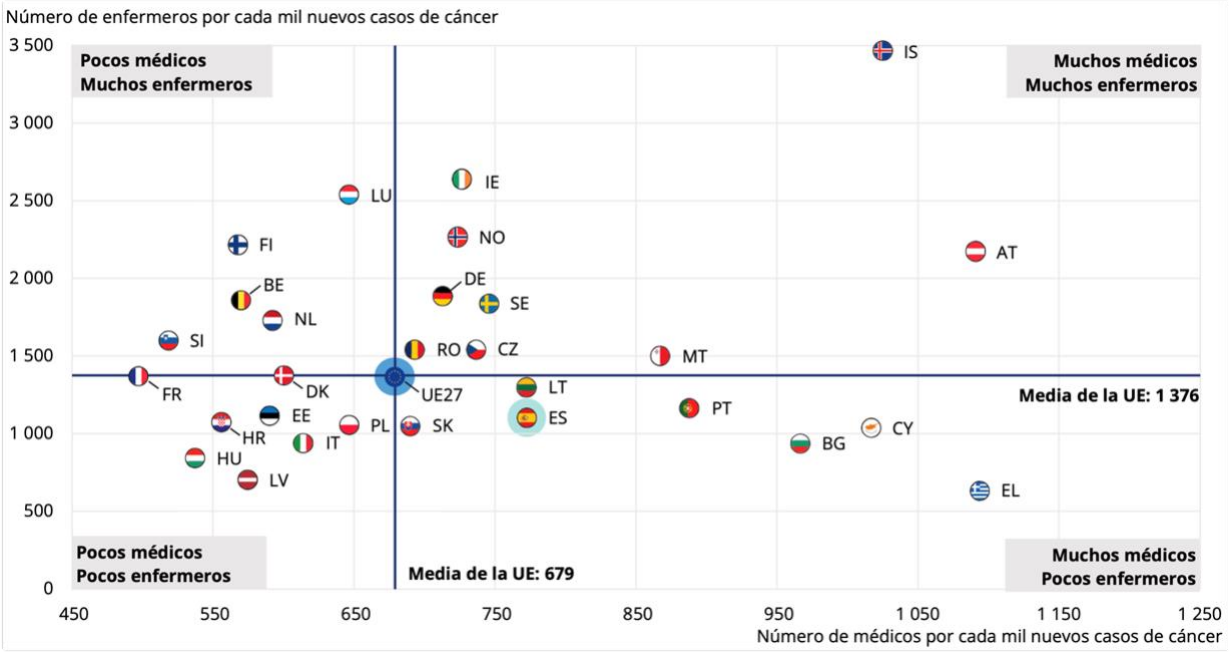


Tabla 6 Distribución de recursos por país

Fuente: Estadísticas sobre salud de la OCDE de 2024.

1.4.2 Costes

En base al informe de Enero del 2020 hecho por Oliver Wyman para la Asociación Española Contra el Cáncer (AECC), en España 1 de cada 2 hombres y 1 de cada 3 mujeres serán diagnosticados de cáncer a lo largo de su vida. Como dato relevante, en el año 2019 se diagnosticaron 275.562 nuevos casos calculando que la enfermedad afecta a 1,5 millones de personas en España.

En total, se estima que el cáncer cuesta a la sociedad española alrededor de 19.300 millones de euros. Para evaluar los costes del cáncer hay que tener en cuenta diferentes variables para cada tipo de paciente, por ejemplo, la edad, género, tipología del cáncer, etc.

El informe de Wyman clasifica los costes en tres grandes tipologías: Costes directos médicos, costes directos no médicos y costes indirectos.

Los costes directos médicos se refieren a todos los costes médicos que tiene una persona por el hecho de tener cáncer, por ejemplo, los tratamientos, seguimientos, costes de farmacia, cuidados paliativos, etc.

Los costes directos no médicos son todos aquellos costes que tiene una persona con cáncer que no están relacionados con la medicina, por ejemplo, el transporte, comida, alojamiento, equipamiento, reacondicionamiento del hogar para adaptarlo por alguna discapacidad del paciente, etc.

Los costes indirectos están relacionados con la pérdida de productividad tanto del paciente como de la familia a causa del cáncer, por ejemplo, los ingresos activos del paciente, ingresos del hogar, etc.

1.4.3 Cálculo de los costes

Los costes de un cáncer pueden variar por diferentes razones: la tipología de cáncer, el estadio del diagnóstico, la edad y el género. También es importante agregar la variable de quién asume el coste.

Basándonos en el estudio de Wyman, se realizan los siguientes cálculos:

El estudio utiliza las tipologías de cáncer que representan aproximadamente el 67% de los casos y calcula el resto como una media de las principales. Dentro de cada tipo de cáncer, se hace una distinción por estadio al diagnóstico, entre local y metastásico.
Las tipologías de cáncer más representativas son: colorrectal, mama, próstata, pulmón, vejiga, hematológico.

Se definen cinco grupos de edad: de 0 a 17 años, de 18 a 44 años, de 45 a 54 años, de 55 a 65 años y mayor o igual a 65 años.

Una vez realizados los cálculos, se define que porcentaje del coste es asumido por el sistema sanitario o por las familias. En el siguiente gráfico se aprecian cómo están distribuidos los costes:

Tabla 7 Gráfico con los totales y los desgloses por apartados

Fuente: Oliver Wyman

Para estimar el cálculo de los costes directos médicos, en los tratamientos, se diferencia entre casos de cáncer local y casos de cáncer metastásico. En el caso de cáncer local, se ha multiplica el número de pacientes con cáncer local (desglosado por tipología) por el coste medio del cáncer por tipología. En estos casos se asume que el cáncer se trata y se cura.

En el caso de cáncer metastásico, se multiplica el número de pacientes con cáncer metastásico (desglosado por tipología) por el coste medio del cáncer por tipología. En estos casos se asume que el cáncer no se cura, por lo que el paciente estará en tratamiento hasta que fallezca.

Para calcular los costes por seguimiento también se diferencian entre los casos de cáncer local y casos de cáncer metastásico.

En cáncer local, se asume una media de 5 años de seguimiento, un parámetro definido a partir de entrevistas con oncólogos. Las especificidades de cada año de seguimiento realizado difieren por tipología de cáncer.

En cáncer metastásico se asume que los pacientes están sujetos al tratamiento principal hasta que fallecen, por lo que utiliza la esperanza de vida para calcular el coste de seguimiento, es decir, multiplica el coste de seguimiento anual por la esperanza de vida en años, excluyendo el primer año de tratamiento (que ya ha sido incluido en el apartado de costes de tratamiento).

En cuanto a los costes de farmacia y parafarmacia, se calculan los pagados por el paciente (gasto medio por las principales tipologías de cáncer).

Por otro lado, están los cuidados paliativos. Son aquellos costes que no van a solucionar el problema del paciente, sino que sirven para acompañarlo y evitándoles el menor sufrimiento posible. Para calcularlos se recogen el consumo de recursos sanitarios en los últimos meses de vida de la población que fallece por cáncer. El coste medio de este tipo de cuidados se divide por tipologías:

Tipología	Coste medio
Colorrectal	10.681 €
Mama	7.832 €
Próstata	10.108 €
Pulmón	12.619 €
Vejiga	12.562 €
Hematológico	17.179 €
Otro	11.178 €

Tabla 8 Coste medio de cuidados paliativos según tipología de cáncer

Fuente: Oliver Wyman

Los costes directos no médicos, como se ha mencionado anteriormente, recogen aquellos que no son médicos en los que depende una persona por el hecho de tener cáncer, como ser: transporte, comida y alojamiento, equipamiento y obras (por ejemplo, obras en el hogar del paciente), cuidados formales, cuidados informales y transporte a radioterapia subsidiado por el Estado. Estos gastos están asumidos mayoritariamente por las familias y suponen aproximadamente el 12% de la estimación total, lo que significa alrededor de 2.220 millones de euros.

Los costes indirectos están relacionados con la pérdida de productividad tanto del paciente como de las familias como consecuencia directa del cáncer; pérdida de ingresos de pacientes activos, pérdida de ingresos del hogar (excluyendo a los pacientes) y pérdida de productividad por muerte prematura.

Se estima que estos costes abarcan aproximadamente el 40% de la estimación total, lo que significa alrededor de 7.700 millones de euros.

El coste se calcula multiplicando la pérdida mensual de ingresos por la tasa de ocupación según género y tramo de edad y por los siguientes parámetros, en función de la casuística de la que se trate:

- Casos de cáncer metastásico: se multiplica además por la esperanza de vida (ya que se asume que todos los pacientes fallecen, pero hasta que fallecen también dejan de generar ingresos) y se descuenta a valores actuales.
- Casos de cáncer local que se curan: se multiplica además por la tasa de supervivencia y los años de vida laboral por delante y se descuenta a valores actuales.
- Casos de cáncer local que fallecen: se multiplica además por la tasa de mortalidad a 5 años para estimar el porcentaje de pacientes de cáncer local que fallecen. En estos casos, se asume que todos los pacientes fallecen, de media, a los 2,5 años (la media entre 0 y 5 años), por lo que se calcula la pérdida de ingresos de dichos pacientes durante 2,5 años, descontado a valores actuales.

La fórmula para hacer estos cálculos es la siguiente:

$$(A * Ts * Sa) + (A * Tf * Sb) + (B * Sc)$$

***A** = Incidencia de cáncer local*
***B** = Incidencia de cáncer metastásico*
***Ts** = Tasa de supervivencia a 5 años*
***Tf** = Tasa de fallecimiento a 5 años*
***Sa** = Salario de por vida descontado a valores actuales*
***Sb** = Salario de la esperanza de vida de 2,5 años de las personas que fallecen en los 5 primeros años desde el diagnóstico descontado a valores actuales.*
***Sc** = Salario de la esperanza de vida según tipología de cáncer y descontado a valores actuales.*

En los siguientes datos se pueden apreciar el monto aproximado del dinero que pierden los pacientes mensualmente estando o no activos en el ámbito laboral:

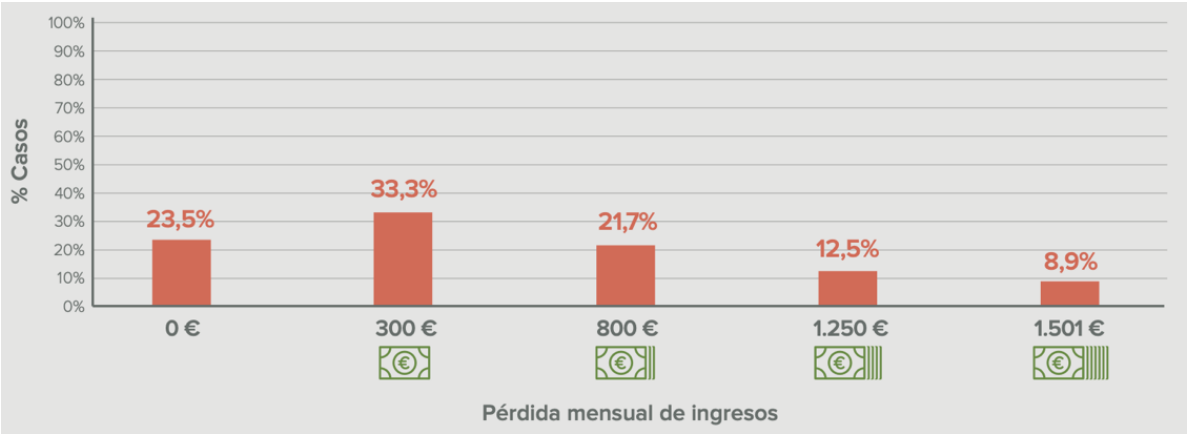


Tabla 9 Pérdida mensual de ingresos de pacientes activos a consecuencia del cáncer

Fuente: Oliver Wyman

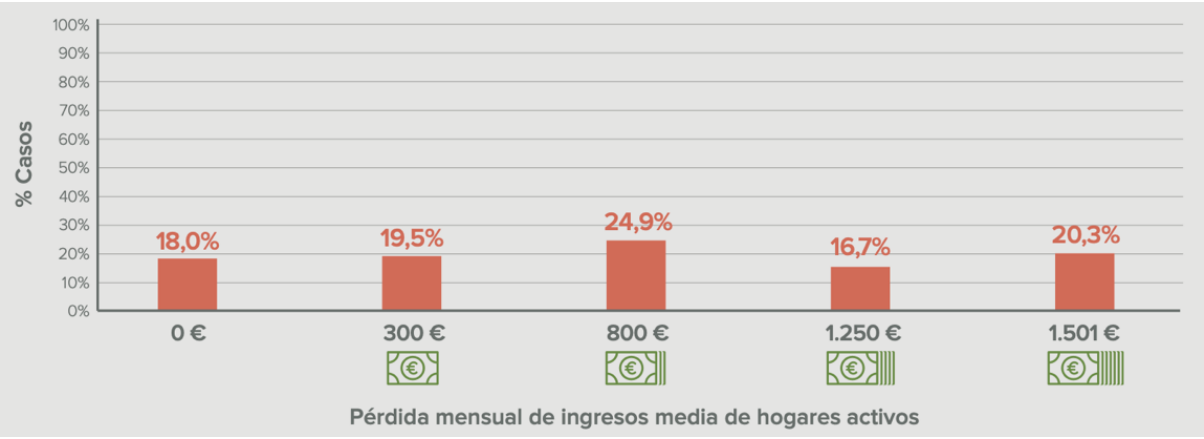


Tabla 10 Pérdida mensual de ingresos de hogares activos a consecuencia del cáncer incluye pacientes

Fuente: Oliver Wyman

Persona afectada	Pérdida de ingresos mensual (en euros)
Hogar + paciente	771
Paciente	564
Hogar (excl. Paciente)	208

Tabla 11 Pérdida media mensual de ingresos, según paciente vs hogares (año 2019)

Fuente: Oliver Wyman

1.5 Objetivos

En esta ocasión como se ha mencionado anteriormente este TFM tiene un doble objetivo que se puede resumir en un objetivo general que se basa en el desarrollo de un modelo basado en técnicas de Business Intelligence que permita mejorarla detección precoz del cáncer de mama y optimizar los recursos hospitalarios asociados, contribuyendo así a una mayor eficiencia en la atención sanitaria y una reducción del gasto médico.

A su vez es posible definir objetivos específicos, estos darán lugar a la composición del índice, donde en primer lugar se definirán cada uno de los términos y de las áreas en las que se profundizara. Se ha resumido en los siguientes objetivos:

En primer lugar, se procederá al diseño de un modelo predictivo de clasificación para la detección temprana del cáncer de mama, aplicando técnicas de análisis de datos y machine learning, con el fin de identificar cuáles son los patrones que favorezcan a la realización de diagnósticos más eficaces y oportunos. Para ello se comparará con los modelos que se conocen en busca de aquel que ofrezca mayor precisión y tasa de acierto, ya que a nivel de salud un mal diagnóstico puede llegar a suponer perder una vida.

En segundo lugar, se busca analizar el impacto económico del cáncer de mama en el sistema sanitario español. Centrando en estudio en los gastos de esta enfermedad se analizará que supone y que impacto tiene sobre los hospitales y también sobre los pacientes.
Los gastos principales que se analizarán son aquellos costes derivados del proceso diagnóstico y del tratamiento, para cuantificar su peso dentro del gasto público total.

En tercer lugar y siguiendo la idea de reducir los costes, se procederá a desarrollar un modelo de regresión orientado a estimar el gasto hospitalario asociado al cáncer de mama, en función de variables clínicas, demográficas y de diagnóstico, que sirva como herramienta para la planificación y gestión de recursos sanitarios.

En cuarto lugar, se implementarán herramientas de Business Intelligence para el análisis y visualización de los datos oncológicos, facilitando la comprensión y la toma de decisiones por parte de profesionales clínicos, gestores y responsables de políticas sanitarias. La finalidad es elaborar un dashboard donde se controle la incidencia del cáncer y como se están gestionando los recursos, de forma que permita a los organismos decisores distribuir de forma correcta y eficiente los recursos tanto monetarios como médicos.

Por ellos finalmente se elaborarán propuestas de mejora para optimizar el uso de recursos sanitarios y promover estrategias de prevención eficaces, basadas en los resultados del análisis de datos, con el fin de mejorar el pronóstico de las pacientes y reducir la carga económica del cáncer de mama en el sistema de salud.

1.6 ODS en la detección del cáncer de mama y la reducción de gastos en la sanidad

Constantemente se oye mencionar la agenda 2030 o los Objetivos de Desarrollo Sostenible (ODS) hay que entender y comprender que relevancia tiene esto en cada una de las acciones y medidas que los gobiernos toman.
Estos objetivos han sido establecidos por las Naciones Unidas en 2015 como parte de la conocida Agenda 2030 y constituyen un marco global, es decir que debe ser cumplido por todos los países miembros, orientado a erradicar la pobreza, proteger el planeta y garantizar el bienestar para todos (Asamblea General de las Naciones Unidas, 2015). Son 17 objetivos interrelacionados que abordan los principales retos sociales, económicos y ambientales que enfrenta la humanidad. En el ámbito de la salud, la igualdad, la innovación y la eficiencia institucional, los ODS proporcionan una hoja de ruta para orientar acciones sostenibles que generen un impacto positivo y duradero.

Durante este Trabajo Fin de Máster *"Cáncer de mama y Business Intelligence: hacia una detección precoz y eficiente"* se pretende que se aborden los ODS, ya que se centra en abordar un problema de salud pública junto con la necesidad de optimizar el uso de recursos sanitarios.
Según la Red Española de Registros de Cáncer (REDECAN, 2024), se estiman más de 35.000 nuevos casos anuales de cáncer de mama en España, con una supervivencia global a 5 años del 82,8 % que supera el 99 % cuando la enfermedad se detecta en fase localizada (REDECAN, 2024). Desde la vertiente económica, diversos análisis del Sistema Nacional de Salud indican que el coste medio de tratamiento por paciente se sitúa en torno a 14.500 € en estadio I, mientras que en estadio IV supera los 65.000 € (Ministerio de Sanidad, 2023), lo que supone más de cuatro veces la inversión inicial y un mayor impacto indirecto en bajas laborales prolongadas y pérdida de productividad.

Se tiene la intención de integrar los ODS ya que se cree que ofrece un marco para cuantificar y reducir estas cargas sanitarias y económicas, al tiempo que proporciona una propuesta de mejorar sostenible, menos costosa pero más accesible para la sociedad. Además, el uso de herramientas de Business Intelligence (BI) se alinea con la meta común de modernizar los sistemas de salud, mejorar la eficiencia, reducir desigualdades y facilitar una toma de decisiones basada y orientada a los datos.

La detección precoz del cáncer de mama, tal como se plantea en este Trabajo Fin de Máster, mantiene una vinculación directa y multidimensional con diversos Objetivos de Desarrollo Sostenible. El ODS 3: Salud y Bienestar constituye la norma principal en la que se basa este trabajo. A través de la meta 3.4, orientada a reducir la mortalidad prematura por enfermedades no transmisibles, y la meta 3.8, que promueve la cobertura sanitaria universal, la detección temprana mediante programas de cribado eficientes cumple un papel decisivo. Se espera que se consiga facilitar la estratificación de riesgo, el seguimiento de la participación en campañas de prevención y la asignación óptima de recursos, incidiendo directamente en la mejora de los indicadores de salud pública y en la sostenibilidad del sistema.

El ODS 5: Igualdad de Género adquiere especial relevancia, puesto que el cáncer de mama afecta mayoritariamente a mujeres. La analítica avanzada de datos permite identificar brechas de acceso a los servicios preventivos por edad, nivel socioeconómico o ubicación geográfica y diseñar

intervenciones con enfoque de género que reduzcan desigualdades estructurales y mejoren la calidad de vida de las pacientes.

Desde la perspectiva económica, el ODS 8: Trabajo Decente y Crecimiento Económico se centra en la reducción de los costes directos e indirectos derivados del tratamiento de estadios avanzados y en la disminución del absentismo laboral asociado. Tal como muestran las estimaciones del Ministerio de Sanidad (2023), la diferencia de costes entre los estadios I y IV supera los 50.000 €, lo que evidencia la relevancia de una intervención temprana para preservar la productividad y liberar recursos públicos.

En el plano tecnológico, el ODS 9: Industria, Innovación e Infraestructura se refleja en la adopción de soluciones de BI, inteligencia artificial y algoritmos predictivos para detectar anomalías en mamografías, priorizar citas y monitorizar resultados clínicos. Estas herramientas refuerzan la infraestructura digital de los sistemas sanitarios y facilitan la investigación aplicada (meta 9.5).

Finalmente, el ODS 17: Alianzas para lograr los Objetivos subraya la necesidad de colaboración entre autoridades sanitarias, universidades, empresas tecnológicas y organismos internacionales. La interoperabilidad de bases de datos, el uso de estándares abiertos y la transparencia en el intercambio de información resultan esenciales para desplegar proyectos de BI de alto impacto y alimentar circuitos de mejora continua.

La detección temprana del cáncer de mama mediante herramientas de Business Intelligence representa una oportunidad real para avanzar en el cumplimiento de varios ODS clave. Este TFM, al centrarse en la salud, la equidad, la eficiencia económica y la innovación tecnológica, se posiciona como una contribución concreta a la Agenda 2030 y como un ejemplo de cómo la analítica de datos puede potenciar la sostenibilidad de los sistemas de salud.

2. Cáncer de mama

2.1. Situación actual del cáncer de mama

2.1.1 Qué es el cáncer de mama

Hablar del cáncer se ha vuelto algo común en nuestra sociedad, no solo por la cobertura mediática o los avances científicos continuos, sino por su alta incidencia en la población. Desgraciadamente, esta enfermedad afecta directa o indirectamente a un gran número de personas en todo el mundo, lo que ha generado una creciente concienciación social y sanitaria en torno a su diagnóstico, tratamiento y prevención.

No obstante, como se ha mencionado anteriormente el término "cáncer" engloba un grupo amplio y heterogéneo de patologías oncológicas que se desarrollan de manera diferente en función del órgano o tejido afectado. Cada tipo de cáncer presenta una sintomatología bastante específica, por lo que responde a tratamientos distintos y muestra tasas de supervivencia diferentes, lo que implica también diferencias significativas en términos de impacto clínico, social, emocional y económico.

Este amplio abanico exige que se lleve a cabo un enfoque individualizado en su estudio y tratamiento, ya que las estrategias de prevención, diagnóstico precoz y atención médica deben adaptarse a las particularidades de cada variante oncológica y a cada paciente. En el caso concreto se ha optado por centrar el análisis en una de las formas más prevalentes y estudiadas: el cáncer de mama.

El cáncer de mama representa la neoplasia maligna más frecuente en mujeres según la doctora Paula Díez Sánchez, aunque también es posible que se de en hombres, a nivel mundial, y su incidencia ha aumentado de forma considerable en las últimas décadas. Aunque los avances médicos han permitido mejorar los índices de detección temprana y aumentar la tasa de supervivencia, sigue siendo una de las principales causas de muerte por cáncer en la población femenina, lo que muestra una necesidad de continuar profundizando en su estudio desde múltiples disciplinas, incluida la aplicación de herramientas tecnológicas como el machine learning.

Según la doctora Ana Santaballa Bertrán, oncóloga médica española, en un informe publicado por la Sociedad Española de Oncología Médica (SEOM), el cáncer de mama se define como “la proliferación acelerada e incontrolada de células del epitelio glandular”. Esta descripción hace referencia al crecimiento anormal y/o acelerado de células malignas o no en los tejidos mamarios, que pueden invadir tejidos cercanos o llegar a otras partes del cuerpo mediante la metástasis. (Sociedad Española de Oncología Médica, 2023.)

Para comprender mejor la enfermedad y así conocer cómo se origina y su posterior evolución, es útil entender como es la estructura anatómica de la mama y donde se produce este crecimiento de células cancerosas. En la siguiente imagen se observa la anatomía de una mama femenina.

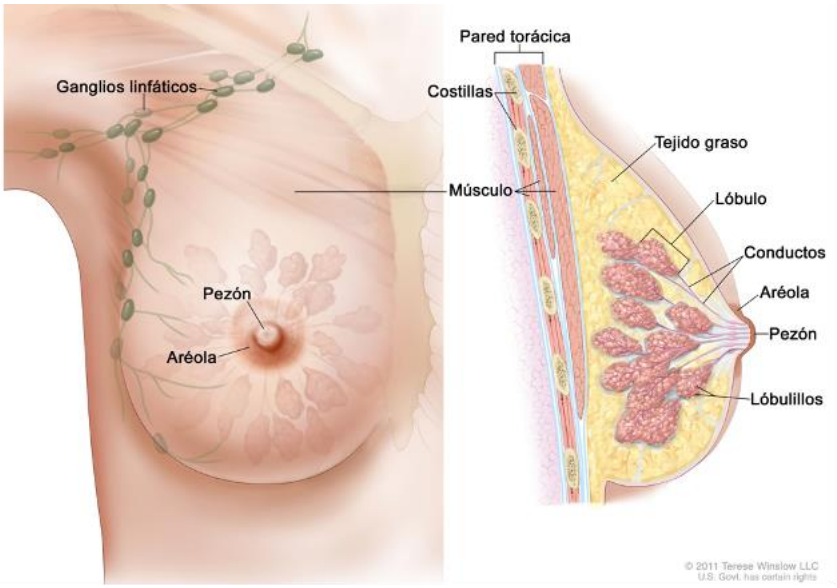


Ilustración 1 Anatomía de la mama femenina

Fuente: SEOM

La glándula mamaria, mama o seno está compuesta por entre 15 o 20 lóbulos, que son secciones glandulares distribuidas alrededor de la mama formando una especie de circunferencia. A su vez, cada lóbulo se subdivide en unidades más pequeñas denominadas lobulillos, que contienen las glándulas productoras de leche durante la lactancia. (Anatomy of the Breasts, n.d.)

Estos lobulillos están conectados por una especie de conductos que reciben el nombre de conductos, estos se encargan de transportar la leche que se ha producido en los lobulillos hasta los pezones. Como se observa en la figura todo ese espacio que hay desde los lobulillos hasta los conductos es tejido graso y fibroso, además de vasos sanguíneos y linfáticos. (Definición De Conducto Galactóforo - Diccionario De Cáncer Del NCI, n.d.)

Es importante hablar de los vasos linfáticos, estos son los encargados de transportar la linfa, un líquido acuoso e incoloro, hasta los ganglios linfáticos que ya se encargan de filtrar la linfa y almacenar glóbulos blancos que defienden y ayudan a combatir al organismo de enfermedades e infecciones. Se puede decir que su función es la de proteger o filtrar, por ello atrapa las bacterias, las células cancerosas o tumorales y cualquier sustancia que pueda resultar nociva. Esto es importante ya que el ser humano no solo posee ganglios linfáticos en las axilas si no que estos se distribuyen por todo el cuerpo. (*Sistema Linfático: MedlinePlus Enciclopedia Médica*, n.d.)

Como afecta a estos ganglios es relevante para entender y comprender en qué punto se encuentra la enfermedad. Uno de los principales síntomas que las personas notan es la “aparición de bultos en la zona de la axila”, estos “bultos” hacen referencia a la afectación de los ganglios por ello la evaluación de estos es fundamental en el diagnóstico y estadificación donde tras su análisis por medio de una biopsia se puede medir en qué punto se encuentra el cáncer, si se sufre de recaída, si se sufre de metástasis o la gravedad/estadio.

En el caso del cáncer de mama, los tumores pueden originarse principalmente en dos estructuras anatómicas: los lobulillos y los conductos. Estas áreas por su función y estructura celular son especialmente susceptibles.

En las siguientes dos imágenes se presenta una comparativa visual entre el estado normal de un lobulillo y un conducto (es decir, sin presencia de cáncer) y cómo se ven cuando existen células cancerosas. La diferencia es apreciable: mientras que en un tejido sano las células están organizadas y mantienen una morfología regular, en el tejido afectado por cáncer se observan un exceso de células con estructura anormal, es decir se pierde de la estructura habitual y aparece una acumulación desordenada de células cancerosas.

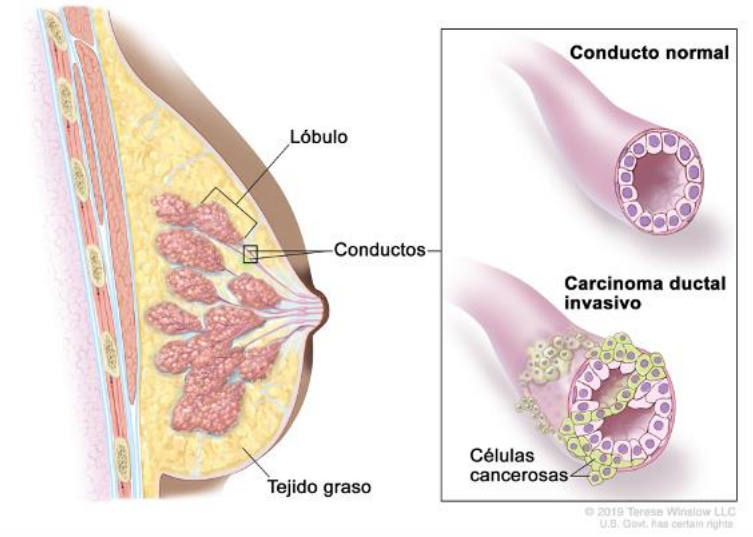


Ilustración 2 Carcomía condutal invasivo de mama

Fuente: NIH

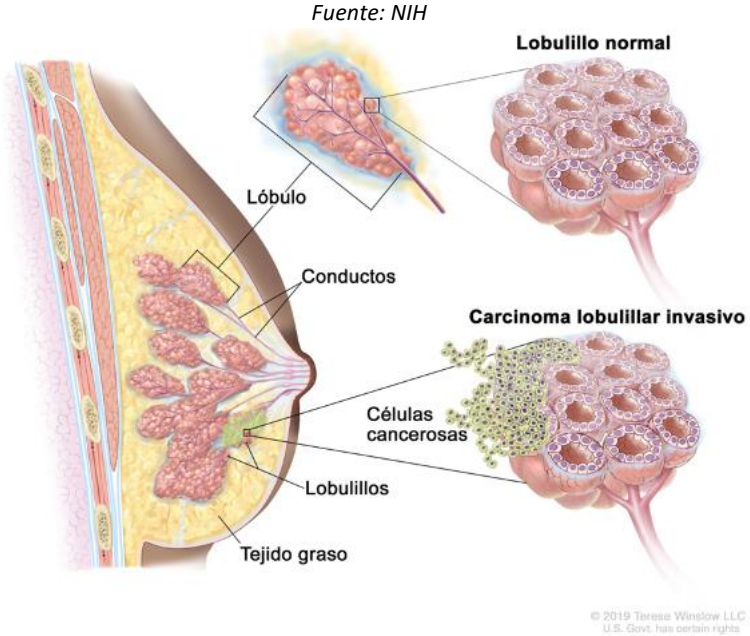


Ilustración 3 Carcomía lobulillar invasivo de mama

Fuente: NIH

Las células cancerosas que se han visualizado en las siguientes imágenes cuando se logra a detectar la enfermedad a tiempo o es demasiado grave llegan a invadir los vasos linfáticos o sanguíneos que provocará que estas células lleguen a otras partes del cuerpo originando la metástasis.

La capacidad de invasión y propagación que tiene el cáncer de mama lo convierte en una enfermedad bastante compleja y potencialmente sistémica, que posee la capacidad de afectar al resto de órganos del cuerpo humano como son los pulmones, el hígado, los huesos o el cerebro.

En ocasiones la sospecha de padecer cáncer de mama se da por una autoexploración o durante las exploraciones clínicas o mamografías que hacen a las mujeres en los centros médicos. Tras detectarlo se inicia un proceso de diagnóstico donde lo primero que se suele realizar son pruebas de imagen y los estudios, como son:

- Examen físico y antecedentes de salud
- Examen clínico de la mama
- Estudios bioquímicos
- Mamografías
- Ecografía
- Resonancias magnéticas

Finalmente, para obtener un diagnóstico definitivo se requiere de la realización de la biopsia, donde se obtiene una muestra del tejido para su análisis.

2.1.2 Tipos de cáncer de mama

Tal como se mencionó anteriormente, el cáncer de mama suele originarse en los lobulillos o en los conductos mamarios. Su estructura puede visualizarse fácilmente a través de esquemas que representan el tejido sano frente al tejido afectado.

Sin embargo, más allá de su localización anatómica, es importante señalar que el cáncer de mama no se presenta como una única enfermedad con síntomas claros, sino que encuentran múltiples subtipos con características clínicas diferentes. Esta diversidad es precisamente uno de los factores que dificulta su detección temprana, ya que algunos subtipos pueden presentar síntomas atípicos o evolucionar de forma más silenciosa, lo que puede llevar a diagnósticos tardíos.

Según la NIH los tipos de cáncer de mama pueden clasificarse según dos criterios principales: la clasificación histológica, basada en la estructura celular y el lugar de origen del tumor, y la clasificación molecular, basada en la expresión de ciertos biomarcadores que permiten guiar los tratamientos y prever la evolución clínica de la enfermedad.

1. Clasificación histológica

Como se ha mencionado en esta clasificación se hace el enfoque en el tejido afectado y el grado de invasión al tejido que se encuentra alrededor. Entre los tipos más comunes se encuentran:

- Carcinoma ductal in situ (CDIS)

Es un tipo de cáncer que se considera a priori no invasivo, este se origina en los conductos mamarios. Suele presentar células cancerosas dentro den conducto sin invadir el tejido graso que las rodea. Aunque puede ser invasivo si se detecta a tiempo posee una tasa alta de supervivencia.

- Carcinoma ductal invasivo (CDI)

Actualmente representa entre el 65% y el 85% de los casos invasivos. Similar al CDIS, pero en este caso las células cancerosas si han invadido el tejido grasoso y los ganglios linfáticos. En función de cómo este de avanzado y cuando se haya detectado puede suponer la supervivencia del paciente, por continuar extendiéndose.

- Carcinoma lobular invasivo (CLI)

Supone aproximadamente el 10% de los casos y en esta ocasión se produce en los lobulillos. Es más complicado de detectar ya que tiende a ser multicéntrico y bilateral. Al igual que el CDI en casos muy graves o con detecciones tardías suele llegar a los ganglios linfáticos y extenderse.

- Carcinoma lobulillar in situ (CLIS)

Las células cancerosas se localizan en los lobulillos y no presentan capacidad invasiva, por lo que no han invadido los tejidos cercanos.

- Carcinoma inflamatorio

Es un tipo muy agresivo, ya que las células cancerosas han bloqueado los vasos linfáticos y causan enrojecimiento, calor e hinchazón. Es complicado de detectar con los métodos convencionales, pues mediante el tacto no suele verse. Suele aparecer en estadios o casos muy graves.

- Enfermedad de Paget del pezón

Afecta la piel del pezón y la areola. Se manifiesta con enrojecimiento, descamación o picor persistente en la zona.

- Tumores filoides y angiosarcomas

Tipos de cáncer raros, menos comunes, pero presentan un comportamiento muy agresivo y con bajas tasas de supervivencia.

2. Clasificación molecular

La clasificación molecular ha supuesto un avance clave para la medicina personalizada, ya que permite adaptar el tratamiento al perfil biológico del tumor. Esta clasificación se realiza en base a la presencia o ausencia de ciertos receptores:

- Luminal A → Implica un buen pronóstico y se trata con terapia hormonal.
- Luminal B → Implica un pronóstico intermedio y se trata con tratamientos combinados.
- HER2 enriquecido→ Se considera uno de los pronósticos malos y con terapias algo más agresivas.
- Basal-like / Triple negativo → Es el tipo más agresivo y con peor pronóstico.

Conocer el estado de los receptores y el tipo de cáncer que se tiene es crucial para seleccionar un tratamiento adecuado.

La estadificación es la clasificación que se hace sobre el punto en el que se encuentra la enfermedad, se clasifican en estadios. Un cáncer de mama en estadio I es un cáncer de mama en una etapa inicial y un estadio IV es un cáncer de mama avanzado que se ha extendido a otras partes del cuerpo.

Estadio 0 o carcinoma in situ:
Carcinoma lobulillar in situ: lesión en la que hay células anómalas en el revestimiento del lobulillo. Raramente se convierte en cáncer invasor pero aumenta el riesgo de padecer cáncer de mama tanto en la mama de la lesión como en la contralateral.
Carcinoma ductal in situ o carcinoma intraductal: lesión en la que hay células anómalas en el revestimiento de un conducto. No es una lesión invasiva pero si se deja evolucionar, puede convertirse en un carcinoma infiltrante o invasor.
Estadio I.
El tumor mide menos de 2 cm y no se ha diseminado fuera de la mama.
Estadio II. Incluye cualquiera de los siguientes:
El tumor mide menos de 2 cm pero ha afectado a ganglios linfáticos de la axila.
El tumor mide de 2 a 5 cm (con o sin diseminación ganglionar axilar).
El tumor mide más de 5 cm pero no ha afectado a los ganglios linfáticos axilares
Estadio III o localmente avanzado. A su vez se divide en:
Estadio IIIA. Incluye los siguientes:
El tumor mide menos de 5cm y se ha diseminado a los ganglios linfáticos axilares de forma palpable o a los ganglios situados detrás del esternón.
El tumor mide más de 5 cm y se ha diseminado a los ganglios linfáticos axilares o a los ganglios situados detrás del esternón .
Estadio IIIB. Es un tumor de cualquier tamaño que afecta a la pared del tórax o a la piel de mama.
Estadio IIIC. Es un tumor de cualquier tamaño con: Afectación de más de 10 ganglios axilares. Afectación de ganglios axilares y de ganglios situados detrás del esternón. Afectación de ganglios situados por debajo o por encima de la clavícula.
Estadio IV
El tumor se ha diseminado a otras partes del cuerpo.

Tabla 12 Niveles de Estadio del cáncer de mama

Fuente: SEOM

2.1.3 Tratamientos de cáncer de mama

El tratamiento del cáncer de mama es un proceso complejo y con un carácter muy individual, ya que se debe adaptar a las características específicas del tumor, el estadio en el que se encuentra, el perfil molecular, el estado general de salud de la paciente e incluso las preferencias.

En España en coste del tratamiento es cubierto por la sanidad pública por lo que se encuentra subvencionado, al igual que todas las visitas médicas que se hayan realizado hasta el momento. Además, la detección precoz juega un papel determinante, ya que permite optar por tratamientos menos agresivos y con mayor probabilidad de éxito, lo que tiene un impacto significativo en el pronóstico, en la calidad de vida y por supuesto en los gastos médicos.

Los tratamientos se dividen en dos grandes grupos en los locales, que actúan sobre el tumor o su zona de origen, y sistémicos, que afectan a todo el organismo.

La cirugía es el principal tratamiento en estadios iniciales, especialmente en aquellos detectados precozmente. Existen diferentes tipos de cirugías:

- Cirugía conservadora de la mama.
- Mastectomía: extirpación completa de la mama.
- Biopsia del ganglio centinela
- Reconstrucción mamaria.

La radioterapia, técnica que se usa posterior a la cirugía con el objetivo de eliminar células residuales y reducir la probabilidad de recurrencia en la zona. Se usa en estadios iniciales donde es posible evitar recaídas y erradicar la enfermedad sin necesidad de tratamientos más agresivos.

La quimioterapia, es la técnica más conocida en casos de cáncer. Se administra según las características del tumor y las necesidades que presente la paciente, ya que puede ser clave para evitar la metástasis. Por ello, es considerada la solución más agresiva cuando el cáncer se detecta en una fase temprana suele evitarse o se administra en ciclos más cortos y menos agresivos.

Por otro lado, existen otros tratamientos como la Terapia hormonal (usada en casos de detención precoz) o terapias dirigidas (usado en casos más específicos y complejos o fases avanzadas).

Finalmente, los cuidados paliativos, aunque no es en sí un tratamiento del cáncer es parte del tratamiento donde se controla los posibles efectos secundarios tanto físicos como emocionales, la atención y seguimiento del paciente como el alivio del dolor, apoyo psicológico, y acompañamiento social. Normalmente es la parte más dura dónde la situación suele ser crítica.

Aunque durante todo este proceso debemos de tener en cuenta que existen sesiones de control y de apoyo donde se monitoriza la evolución del paciente en todo momento.

2.1.4 Tasas y Cifras

En España, el cáncer de mama continúa siendo una de las enfermedades oncológicas con mayor incidencia y prevalencia. Según datos de la Sociedad Española de Oncología Médica (SEOM) y la Red Española de Registros de Cáncer (REDECAN), en sus informes de 2024, se diagnosticaron 36.395 nuevos casos de cáncer de mama, lo que refleja una importante carga para el sistema sanitario y para la sociedad en general. Mientras que para 2025 se prevé un incremento, alcanzando los 37.682 casos, esto significa que aproximadamente 1 de cada 8 mujeres españolas desarrollará esta enfermedad a lo largo de su vida según detalla la AECC. (Sociedad Española de Oncología Médica, 2025)

Además, el Observatorio del Cáncer de la Asociación Española Contra el Cáncer (AECC) estima que la prevalencia del cáncer de mama en mujeres españolas alcanza las 151.945 pacientes, evidenciando no solo su elevada frecuencia, sino también que muchas mujeres conviven tras el diagnóstico.

El cáncer de mama sigue siendo una de las principales causas de mortalidad por cáncer entre las mujeres. Según los datos más recientes de la AECC se estima que el número de muertes por cáncer de mama en la población femenina en 2023 fue de 6.759. Esto representa cerca del 15% de todos los fallecimientos por cáncer en la mujer en nuestro país. Aunque si observamos las incidencias en los

últimos 20 años según los datos del INE la mortalidad por cáncer de mama ha descendido en los últimos años esto es debido a los programas de cribado, la mejora de los tratamientos y la detección precoz que se está realizando según SEOM y la AECC La tasa de supervivencia a cinco años vista se sitúa en torno al 85%, esta es una de las más altas entre los distintos tipos de cáncer.

Es complicado indicar un promedio de tiempo que puede durar la enfermedad, analizando los casos y la información facilitada por las asociaciones contra el cáncer se observa que ese tiempo puede variar considerablemente según diversos factores clínicos y personas como pueden ser el tipo de tumor, el estadio en el momento del diagnóstico, la respuesta al tratamiento y el estado general de salud de la paciente.

Hay casos leves de cáncer que con el tratamiento pueden llegar a erradicarse entre 6 meses o un año o en el peor de los casos puede incluso prolongarse entre 5 y 10 años. El problema aparece cuando el cáncer es metastático o de estadio IV pues actualmente no presenta una cura, pero se trata de forma que muchas pacientes logran vivir varios años más con la enfermedad relativamente controlada.

Estas cifras refuerzan la necesidad de seguir impulsando estrategias eficaces de prevención, diagnóstico precoz, atención integral y acompañamiento continuo, tanto desde el ámbito sanitario como desde el social y emocional.

2.2 Gastos asociados al cáncer de mama

2.2.1 Gastos en la vida del paciente

El cáncer, y en particular el cáncer de mama no constituye una enfermedad simple ni uniforme. Su evolución puede variar significativamente de una persona a otra, incluso en casos con características clínicas aparentemente similares.

Como es evidente afrontar un diagnóstico de cáncer de mama supone un reto muy complejo que afecta profundamente a múltiples áreas de la vida de la persona afectada, en este caso especialmente las mujeres. En los casos del cáncer la vivencia de la enfermedad no se ve afectada exclusivamente en el plano sanitario de la persona, si no que como la propia enfermedad comienza a atacar a todas las áreas de la persona. El diagnostico acaba suponiendo una transformación social, laboral y evidentemente económica.

Los gastos que se asocian a esta enfermedad nacen desde el momento en que se sospecha la existencia de un tumor hasta la confirmación, donde se pueden realizar diversas consultas médicas, tratamientos y pruebas para evidenciar la existencia del tumor y su estadio. Posteriormente se debe iniciar el tratamiento y realizar un seguimiento de cómo está afectando en la salud del paciente, toda esta situación lleva a una intensa sensación de incertidumbre y miedo tanto del pronóstico como de si habrá efectos secundarios y como todo afectará su día a día.

A nivel sanitario en España el acceso a las pruebas, los tratamientos y el seguimiento está garantizado gracias a las coberturas que dispone la seguridad social (el sistema nacional de salud) aunque también existe la posibilidad de costearlo mediante la sanidad privada.

Aun así y habiendo analizado anteriormente el impacto que tiene al gasto público, el sistema no absorbe todos los costes que genera la enfermedad. Sino que el sistema cubre gran parte de los gastos directos el resto dependen de los pacientes. Estos se pueden catalogar en gastos directos no cubiertos (como productos de apoyo, prótesis externas, fármacos, tratamientos complementarios o transporte

sanitario), así como costes indirectos asociados, por ejemplo, a la pérdida de ingresos por incapacidad laboral, la reducción de jornada o la necesidad de apoyo domiciliario. Todos estos gastos que deben asumir los pacientes se denomina carga financiera objetiva, y hace referencia al impacto económico cuantificable que la enfermedad genera en las pacientes y sus familias.

La asociación del cáncer española en uno de sus informes sobre los gastos médicos asociados a esta enfermedad lo denomina ***"la toxicidad financiera"*** y lo define de la siguiente manera: *"el término toxicidad financiera del cáncer engloba el conjunto de problemas económicos y dificultades laborales que sufren las personas con cáncer y sus familias. Se trata de un fenómeno con muchas caras, ya que el cáncer implica un aumento de los gastos del hogar (médicos, de farmacia, pequeños equipamientos…) pero, además, el cáncer puede producir una bajada de los ingresos de la persona afectada y de las personas que la cuidan y/o la acompañan."*

Para entender cuáles son los gastos que deben soportar los pacientes se va a analiza una encuesta telefónica de aproximadamente 20 minutos que hizo la Asociación del cáncer española a 615 pacientes de cáncer de mama, en este caso por la exploración de los datos podemos extraer información muy relevante.

Como es evidente el aspecto más notable es el relacionado con los gastos directos, es decir aquellos gastos que se pueden relacionar directamente con la enfermedad, pero hay que entender las dimensiones de esos gastos ya que como se mencionó anteriormente, aunque la seguridad social subvenciona gran parte de estos gastos hay muchos que son necesario para la salud del paciente, pero quedan excluidos de la subvención de la seguridad social.

Una pregunta que puede venir a la cabeza, ¿y si lo subvenciona la seguridad social en que gastan el dinero?

A pesar de que la sanidad pública ofrece las visitas médicas, las pruebas, las terapias, las operaciones e incluso sesiones de rehabilitación según las estadísticas generales del cáncer de mama, que se observan en el Observatorio contra el cáncer de mama, 4 de cada 10 pacientes recurren a las clínicas privadas debido a que los plazos de atención son elevados, las citas se dilatan en el tiempo, necesitan una segunda opinión o por confianza.

Si se observa la encuesta que se realiza en el informe de toxicidad financiera aquellas que recurrieron exclusivamente al privado sus gastos medios en solo gastos médicos ascienden a los 2.000 euros, mientras que si observamos la muestra completa hablamos de unos 722 euros de promedio. Aun así, aunque esos gastos cuando se va por la seguridad social no lo incurran las familias debemos pensar que lo desembolsa el estado.

Otros gastos relevantes son los servicios médicos auxiliares como los fisioterapeutas, psicólogos, nutricionistas, lo que podemos llamar servicios de rehabilitación o atención. Al igual que los demás gastos la sanidad pública ofrece gratuitamente alguno de estos servicios (aunque no todos), pero no siempre con la amplitud, necesidad o rapidez que necesitan las pacientes, en este caso se estima que desembolsan según el informe de toxicidad financiera unos 1.131 euros.

Finalmente, en los gastos médicos se pueden incluir aquellos de transporte, dietas o alojamiento que algunas familias deben hacer, en este caso se observa que el gasto promedio es de 1.737 euros.

En la siguiente tabla se observan los promedios y el porcentaje de familias/pacientes que han incurrido en los gastos anteriormente mencionado.

GASTO	>5.000€	1.001€ A 5.000€	201€ A 1.000€	1€ A 200€	SIN GASTO	PROMEDIO DE GASTO
SERVICIOS MÉDICOS, PRUEBAS DIAGNÓSTICAS Y HOSPITALIZACIÓN	3,7%	8,5%	10,2%	14,7%	62,9%	722€
FISIOTERAPEÚTAS, MUTRICIONISTAS, PSICÓLOGOS, AUXILIARES	4,9%	8,5%	10%	7,4%	69,2%	1.131€
DESPLAZAMIENTOS Y DIETAS POR TRATAMIENTOS	6,9%	42,5%	31%	12,1%	7,5%	1.737€
TOTAL GASTOS MÉDICOS	19,8%	42,2%	25,2%	9%	3,8%	3.590€

Tabla 13 % de familias que incurren en gastos médicos e importe

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Uno de los gastos relevantes durante la lucha contra el cáncer es el relacionado con la adquisición de productos en farmacias y parafarmacias. Actualmente la seguridad social cubre una parte importante de los medicamentos necesarios, aunque no todos los fármacos están completamente subvencionados, sino que existe un grado de aportación de los pacientes o en algunos casos no se encuentra subvencionado y depende del paciente directamente. El grado de aportación del paciente al coste de los medicamentos financiados depende de varios factores, como la edad, la renta anual, la situación laboral, el grado de discapacidad o si se trata de una persona pensionista.

Cuando se piensa en que incurren en estos gastos se debe pensar que va más allá del coste de unas simples pastillas. Las pacientes que sufren esta enfermedad deben incurrir en gastos como protectores de estómago, suplementos nutricionales, cosmética específica para pieles sensibles o afectadas por la quimioterapia y radioterapia, cremas para el cuidado de la piel, entre otros. El gasto promedio en estos casos es de 2.283 euros

GASTOS EN FARMACIA Y PARAFARMACIA	>5.000€	1.001€ A 5.000€	201€ A 1.000€	1€ A 200€	SIN GASTO
% DE FAMILIAS QUE HAN INCURRIDO EN GASTOS FARMACÉUTICOS DESGLOSADOS POR ESCALÓN DE GASTO	6%	35,4%	37,9%	13,9%	1,4%

PROMEDIO DE GASTO 2.283€

Tabla 14 % de familias que incurren en gastos farmacéuticos

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

El tratamiento médico genera un gran número de efectos secundarios, si se os dijese de pensar en un efecto visible en los pacientes con cáncer de mama lo primero que viene a la cabeza es la pérdida del pelo y la pérdida de las mamas o parte de ella. Aunque esto va más allá, pues la vida de los pacientes con cáncer de mama cambia drásticamente y se debe recurrir a gastos que ayudan a facilitar y mejorar la vida diaria, en este caso se habla de los productos ortoprotésicos e incluso obras de adaptación.

La enfermedad lleva consigo un efecto psicológico muy grande que la paciente debe asimilar prácticamente de la noche a la mañana, parte de estos productos les ayuda a recobrar cierta normalidad que se considera psicológicamente indispensable como son las pelucas, los sujetadores con prótesis mamarias. Como es lógico en estos gastos influye más que en otros la renta de la que disponen los pacientes, aun así, se estima que el gasto promedio es de aproximadamente 900 euros.

GASTO	>5.000€	1.001€ A 5.000€	201€ A 1.000€	1€ A 200€	SIN GASTO	PROMEDIO DE GASTO
GASTO EN PORDUCTOS ORTOPROTÉSICOS Y AUXILIARES	0%	0%	71,4%	18,2%	10,4%	302€
GASTO EN OBRAS Y GRANDES EQUIPAMIENTOS	1,9%	12%	3,8%	0,7%	81,6%	594€
TOTAL GASTOS OBRAS Y EQUIPAMIENTO	3,6%	11,3%	60,4%	15,5%	9,2%	896€

Tabla 15 % de familias que incurren en gastos para facilitar la vida

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

En algunos casos muchas familias deben recurrir a contratar personal para el cuidado del paciente o para el cuidado doméstico. El coste en este caso depende del poder adquisitivo de las familias, aunque para aquellas familias con un poder adquisitivo muy bajos existen ayudas para costear la asistencia. Aun así, el gasto promedio es de unos 2.500 euros aproximadamente.

PROMEDIO DE GASTO 2.473€

GASTO EN CUIDADOS A LA ENFERMA Y AYUDA EN TARAS DEL HOGAR	>5.000€	1.001€ A 5.000€	201€ A 1.000€	1€ A 200€	SIN GASTO
% DE FAMILIAS QUE HAN INCURRIDO EN GASTOS FARMACÉUTICOS DESGLOSADOS POR ESCALÓN DE GASTO	10%	13,1%	5%	0,5%	71,4%

Tabla 16 % de familias que incurren en gastos de cuidado

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Como se observa mantener la vida con normalidad se vuelve muy complejo para los pacientes con cáncer de mama. Más allá de lo material desde el minuto uno de la detección nace un sentimiento de angustia, miedo y ansiedad a la que se le comienza a añadir un sentimiento de angustia financiera

subjetiva. Existe un sentimiento de preocupación constante relacionado con la sostenibilidad económica futura, el temor a no poder afrontar gastos esenciales o a convertirse en una carga para el entorno.

Poco a poco se observa como comienzan a dejar de poder hacer cosas que antes no requerían esfuerzo, se encuentran más débiles, náuseas y molestias constantes que en la mayoría de los casos afecta a la vida laboral de los pacientes e incluse de su entorno más cercano. Según el informe del Observatorio contra el cáncer de mama dos de cada tres hogares acaban perdiendo ingresos económicos y es que muchas pacientes deben ausentarse temporalmente del trabajo (excedencias o bajas) o incluso muchas acaban perdiendo sus empleos.

PÉRDIDA DE INGRESOS POR CULPA DEL CÁNCER DE MAMA	MÁS DE 50.000€	25.001€ A 50.000€	10.001€ A 25.000€	1€ A 10.000€	SIN PÉRDIDAS
% FAMILIAS INCLUIDAS EN CADA ESCALÓN DE PÉRDIDA DE INGRESOS	16,4%	9,2%	11,1%	26,6%	36,7%

PROMEDIO DE PÉRDIDA DE INGRESOS 32.587€

Tabla 17 % de familias que incurren en pérdida de ingresos

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Además, cuando no es posible depender de personal externo de pago para el cuidado y mantenimiento del hogar este gasto se traduce en un gasto oculto que la familia debe llevar, ya bien sea por reducción de horas o dependencias de otros miembros.

% PACIENTES / HOGARES INCLUIDAS EN CADA ESCALÓN DE PÉRDIDA DE INGRESOS	-67% A -100% INGRESOS	-34% A -66% INGRESOS	-1 A -33% INGRESOS	NO HAN PERDIDO INGRESOS
PACIENTE	28,7%	14,8%	10,9%	45,6%
HOGAR	16,2%	23,4%	23,8%	36,6%

Tabla 18 % ingreso perdido

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Para intentar paliar todos estos gastos y con parte de ayuda de lo subvencionado por la seguridad social es posible recurrir a una serie de pensiones, seguros y ayudas que pueden recibir las familias para reducir todos estos costes. La mayoría vienen dadas por el estado, aunque también es posible que vengan de ONG o algunos seguros.

El importe recibido por las familias que viene por parte del estado y de las ONG es estudiado anteriormente teniendo en cuenta varios motivos de su vida y situación económica, aún así se estimas que el promedio recibido por las familias es de 17.500 euros, teniendo en cuenta que habrá familias que no reciban nada y otra que puedan costearse gran parte de la enfermedad con estas ayudas.

En la siguiente tabla se indicada un resumen de todos los gastos desglosados para entender a qué dimensión pertenecen.

GASTOS PAGADOS POR LAS FAMILIAS	
GASTOS MÉDICOS	• **ATENCIÓN MÉDICA PRIVADA PAGADA POR LA PACIENTE:** CONSULTAS A MÉDICOS, PRUEBAS DIAGNÓSTICAS, HOSPITALIZACIÓN…. • **OTROS SERVICIOS MÉDICOS PRIVADOS:** REHABILITADORES, FISIOS, NUTRICIONISTAS, PSICÓLOGOS… • COSTE DE **TRANSPORTE Y DIETAS** DE LA PACIENTE Y ACOMPAÑANTES PARA LLEGAR AL CENTRO DONDE RECIBE TRATAMIENTO.
GASTOS FARMACÉUTICOS	• **MEDICAMENTOS:** COPAGOS Y RECETAS PRIVADAS • **PARAFARMACIA:** BATIDOS, PROBIÓTICOS, CREMAS, PROTECTOR SOLAR…
OBRAS Y EQUIPAMIENTO	• **PRODUCTOS ORTOPROTÉSICOS Y AUXILIARES:** PELUCA, SUJETADOR, PRÓTESIS MAMARIA, MANGUITO PARA LINFEDEMA… • **OBRAS DE ACONDICIONAMIENTO DEL HOGAR.** • **EQUIPAMIENTO:** CAMAS ADAPTADAS, SILLA DE RUEDAS…
GASTOS DE ATENCIÓN A LA ENFERMA Y TAREAS DEL HOGAR	• **PERSONAL CONTRATADO PARA CUIDAR DE LA PACIENTE** QUE PIERDE AUTONOMÍA POR LA ENFERMEDAD. • **PERSONAL CONTRATADO PARA EXTERNALIZAR TRABAJOS DEL HOGAR** QUE ANTES REALIZABA LA PACIENTE
GASTOS OCULTOS	
CUIDADOS DE FAMILIA Y AMIGOS	• **TRABAJO EXTRA NO REMUNERADO** DE OTROS MIEMBROS DEL HOGAR PARA CUIDAR O SUSTITUIR A LA PACIENTE EN TAREAS DEL HOGAR QUE REALIZABA CON ANTERIORIDAD.
INGRESOS PERDIDOS	
PÉRDIDA DE INGRESOS POR CULPA DEL CÁNCER	• **PÉRDIDA DE INGRESOS DE LA PACIENTE:** DESPIDOS, PÉRDIDA DE ACTIVIDAD, REDUCCIÓN DE HORARIO … • **PÉRDIDA INGRESOS DEL CUIDADOR** DESPIDOS, PÉRDIDA DE ACTIVIDAD, HORAS PERDIDAS, REDUCCIÓN DE HORARIO …
INGRESOS CONSEGUIDOS	
AYUDAS, PENSIONES Y SEGUROS	• APORTES ECONÓMICOS DE **FAMILIA Y ALLEGADOS.** • AYUDAS ECONÓMICAS DE **ONGS.** • **PENSIONES** POR ENFERMEDAD O INCAPACIDAD LABORAL. • INDEMNIZACIONES DE **SEGUROS PRIVADOS** EN FORMA DE RENTA O CAPITAL.

Tabla 19 Resumen de los gatos asociados al cáncerç

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Como se observa medir la dimensión que tiene el cáncer de mama es bastante complicado y personal pero lo que sí se puede observar es que va más allá de una condición médica. Impacta notablemente en la situación económica de los que la sufren y sus familias que acaba alargándose tras haber superado la enfermedad.

GASTOS	>50.000€	25.001€ A 50.000€	5.001€ A 25.000€	1€ A 5.000€	SIN GASTO	PROMEDIO DE GASTO
MÉDICOS	0,3%	1,2%	18,4%	76,4%	3,7%	3.590€
FARMACIA Y PARAFARMACIA	0,1%	0%	11,3%	87,3%	1%	2.283€
OBRAS Y EQUIPAMIENTO	0%	0%	3,6%	87,2%	9,2%	896€
CUIDADOS Y AYUDA EN EL HOGAR	1,1%	1,3%	7,6%	18,6%	71,4%	2.473€
TOTAL GASTOS PAGADOS	2,2%	6,9%	37,1%	53,8%	0%	9.242€

PÉRDIDAS	>50.000€	25.001€ A 50.000€	5.001€ A 25.000€	1€ A 5.000€	SIN GASTO	PROMEDIO DE GASTO
PÉRDIDA INGRESOS	16,4%	9,2%	15%	22,7%	36,7%	32.587€
TOXICIDAD FINANCIERA VISIBLE	19,2%	14,7%	26,5%	37,8%	1,8%	41.834 €

AYUDAS	>50.000€	25.001€ A 50.000€	5.001€ A 25.000€	1€ A 5.000€	SIN GASTO	PROMEDIO DE GASTO
AYUDAS NO PAGADAS (FAMILIA)	9,9%	4,5%	26,7%	32,1%	26,8%	21.593 €
TOXICIDAD FINANCIERA REAL	29,8%	16%	33,9%	18,9%	1,4%	63.427 €

AYUDAS	>5.000€	1.001€ A 5.000€	201€ A 1.000€	1€ A 200€	SIN GASTO	PROMEDIO DE GASTO
AYUDAS, PENSIONES Y SEGUROS	12%	2%	8%	9%	69%	17.536 €

Tabla 20 Resumen monetario de los gatos asociados al cáncer

Fuente: Observatorio del cáncer 2020. Informe de toxicidad financiera.

Teniendo en cuenta que solo se está mencionando el impacto monetario de la enfermedad no puede perderse de vista el daño y sufrimiento oculto que tiene para cada hogar esta situación.

Es relevante recalcar que estos costes varían en función del tipo de cáncer de mama que tenga los pacientes y en que estadio se encuentre. En fases más avanzadas, los tratamientos suelen ser más prolongados, agresivos y conllevan mayores complicaciones físicas, psicológicas y económicas. Por todo ello, la detección precoz no solo mejora significativamente el pronóstico, sino que también puede reducir el coste personal, familiar y social de la enfermedad, permitiendo un abordaje más eficaz, menos invasivo y, en muchos casos, menos costoso.

3. Machine Learning

El aprendizaje automático es una disciplina que tiene como base el desarrollo de modelos estadísticos y algoritmos con el objetivo de permitir a los sistemas informáticos realizar tareas sin necesidad de ser programados explícitamente. Esto es posible conseguirlo sacando deducciones de los datos y encontrando patrones.

El machine learning es una rama de la Inteligencia Artificial, y como su propio nombre indica consiste en que los sistemas aprendan por sí mismas. Es por eso, que como se comenta anteriormente, de acuerdo con los datos proporcionados se utilizan algoritmos para poder analizarlos y mejorar su rendimiento con el tiempo.

En el desarrollo de un modelo de Machine Learning existen cuatro fases. La primera de ellas consiste en elegir la información a utilizar para el entrenamiento del modelo. Esos datos le proporcionaran la capacidad para hallar la solución del problema inicial. Estos se pueden etiquetar para indicarle al modelo las características que debe identificar o en el caso de que se encuentren sin etiquetar, entonces será el modelo el que deberá detectar y extraer características recurrentes. No obstante, etiquetados o no, los datos deben prepararse, organizarse y limpiarse cuidadosamente. De lo contrario, el entrenamiento del modelo de Machine Learning puede estar sesgado y por tanto, los resultados de sus predicciones futuras se verán afectados directamente.

En la siguiente fase, dependiendo del tipo de datos y de su volumen y sobre todo qué tipo de problema se pretende resolver, se selecciona un algoritmo para ejecutar sobre el conjunto de datos de entrenamiento.

Una vez elegido el algoritmo se procede a entrenarlo. En este proceso se ejecutan las variables a y se comparan los resultados obtenidos con los valores reales que debería haber producido. Existen diferentes formas para aumentar la precisión de los resultados, como puede ser ajustar el sesgo o los pesos.

Para finalizar, se valora el uso y la mejora del modelo. Para ello se utilizan nuevos datos, con el fin de evaluar cómo funciona cuando los datos son diferentes. Estos se escogeran teniendo en cuenta el problema a resolver.

En el sector sanitario, el uso y el desarrollo de modelos de machine learning cada vez es más recurrente, ya que ayuda a optimizar recursos, reducir costes y mejorar la calidad de la atención médica utilizando para ello tanto datos clínicos como administrativos.

3.1 Machine Learning supervisado

Según el objetivo del problema a resolver y el conocimiento de los datos a obtener, podemos diferenciar tres tipos de modelos de aprendizaje automático. En el caso de disponer de datos que incluyen tanto la variable objetivo como el resto de las variables independientes, se trata de machine learning supervisado.

Como en todo proceso de modelado de datos, el primer paso consiste en la recopilación y el posterior procesamiento de estos. Esta información puede provenir de diferentes lugares de origen como bases

de datos, registros de eventos, sensores, APIs, etc. La fase del procesado previa a la construcción del modelo tiene como fin garantizar la consistencia y sobretodo la calidad de los resultados. Tras la limpieza de estos datos, cada elemento de entrada recibe una etiqueta correspondiente, las cuales influyen de forma directa en la capacidad de aprendizaje del modelo y en sus predicciones. Por tanto, el etiquetado es fundamental en el machine learning supervisado ya que es necesario enseñar los patrones del modelo para poder predecir.

Durante el entrenamiento, el algoritmo analiza los datos de entrada y aprende a mapearlos a las etiquetas de salida correctas. Este proceso implica ajustar los parámetros del modelo para minimizar la diferencia entre los resultados previstos y las etiquetas reales. El modelo mejora su precisión al aprender de los errores que comete durante el entrenamiento. Una vez que se entrena el modelo, se somete a evaluación. Se utilizan datos de validación para determinar la precisión de un modelo. Dependiendo de los resultados, se ajusta según sea necesario.

En teoría, cuantos más datos absorbe un modelo, más patrones aprende y más precisas se vuelven sus predicciones. El aprendizaje continuo es fundamental en el machine learning: el rendimiento de los modelos mejora a medida que siguen aprendiendo de sets de datos etiquetados.

Este enfoque es especialmente útil cuando se desea predecir un resultado concreto, como el riesgo de reingreso hospitalario o la probabilidad de complicaciones clínicas.

Una vez que se despliega, el machine learning supervisado puede realizar dos tipos de tareas: clasificación y regresión.

3.1.1 Modelos de Clasificación

La clasificación se refiere al problema de garantizar que los algoritmos asignen correctamente una etiqueta de clase a sus conjuntos de datos. Por ello, los algoritmos de clasificación deben ser entrenados por etiquetadores de datos para garantizar que el software de aprendizaje supervisado categoriza su entrada en función de determinados criterios. Además, los modelos de clasificación permiten a los ordenadores identificar a qué categoría pertenece un conjunto de datos, a partir de patrones y características.

Existen muchos modelos de clasificación diferentes, cada uno tienes sus características y fortalezas y funciona de manera diferente, pero todos tienen un objetivo común, el ser capaces de hacer una buena predicción sobre los datos que no ha visto antes. El éxito de la clasificación depende del conjunto de datos y del modelo utilizado. Por ello, existen algunas formas de mejorar la precisión incluyen el preprocesamiento de los datos, el ajuste de hiperparámetros y la selección de características adecuadas.

El Machine Learning de clasificación se utiliza en una variedad de aplicaciones, desde la detección de fraude y la exploración de la genética, hasta el aprendizaje automático en el campo de la inteligencia artificial. Además, se ha convertido en una herramienta de gran valor dentro del ámbito sanitario debido a su capacidad para procesar y analizar grandes volúmenes de datos médicos de manera eficiente y con alta precisión. En un sistema de salud donde cada decisión puede tener un impacto

crítico sobre la vida del paciente, la posibilidad de automatizar ciertos procesos y apoyar a los profesionales mediante modelos predictivos representa una evolución significativa.

Una de las principales razones por las que este tipo de inteligencia artificial resulta útil es que los hospitales y centros de salud generan constantemente una enorme cantidad de información, desde imágenes médicas y resultados de análisis hasta historiales clínicos y notas de los médicos. El ML de clasificación permite identificar patrones en esos datos que pueden facilitar el diagnóstico temprano de enfermedades, clasificar a los pacientes según niveles de riesgo o incluso predecir la evolución de ciertas patologías.

Estos sistemas también se utilizan en contextos de urgencias, donde permiten priorizar la atención de pacientes según la gravedad de los síntomas reportados, lo cual es especialmente útil en hospitales saturados. Además, mediante el análisis automático de textos clínicos, como los informes médicos, es posible clasificar síntomas o diagnósticos, lo que ayuda a mantener los registros electrónicos organizados y actualizados sin una carga administrativa excesiva para el personal de salud.

Desde el punto de vista económico, la implementación de modelos de clasificación también tiene un impacto significativo en la reducción de gastos sanitarios. En primer lugar, permite una optimización de recursos, ya que ayuda a identificar qué pacientes requieren atención prioritaria o tratamientos específicos, evitando intervenciones innecesarias. Esto reduce tanto el uso ineficiente de equipos médicos como la ocupación de camas hospitalarias por pacientes que podrían recibir tratamiento ambulatorio.

En segundo lugar, al facilitar el diagnóstico temprano, objetivo de este trabajo, muchas enfermedades pueden ser tratadas en etapas iniciales, lo que reduce considerablemente los costos asociados a complicaciones posteriores, hospitalizaciones prolongadas o tratamientos más invasivos. Asimismo, los modelos predictivos que anticipan reingresos o complicaciones permiten establecer planes preventivos más baratos y efectivos que el tratamiento reactivo de emergencias.

Por último, el uso del *Machine Learning* puede contribuir a reducir los errores médicos, que son una causa importante de costos tanto humanos como económicos. Al actuar como un sistema de apoyo a la decisión clínica, estos modelos disminuyen la probabilidad de diagnósticos incorrectos o tratamientos inadecuados, lo cual no solo protege la salud del paciente, sino que también evita demandas legales y gastos innecesarios para el sistema sanitario.

A continuación, se presentan y analizan cuatro de los métodos de clasificación más utilizados: Regresión Logística, Árbol de Decisión, Random Forest y Naive Bayes.

La regresión logística es uno de los métodos de clasificación más conocidos y utilizados, especialmente cuando el objetivo es predecir una variable categórica binaria, es decir, con dos posibles valores, como "presencia" o "ausencia" de una enfermedad.

Este modelo utiliza una función logística, también conocida como función sigmoide, para estimar la probabilidad de que una observación pertenezca a una clase determinada. A partir de esta probabilidad, se establece un umbral (comúnmente 0.5) para decidir la clasificación final.

Una de las principales ventajas de la regresión logística es su simplicidad e interpretabilidad. El modelo permite identificar qué variables influyen en la predicción y en qué medida. Esto resulta muy útil en el ámbito clínico, donde es necesario comprender el impacto de cada factor de riesgo. Sin

embargo, este algoritmo supone que las variables predictoras tienen una relación lineal con el logaritmo de las probabilidades, lo cual no siempre se cumple. Además, no se adapta bien a relaciones complejas no lineales ni a interacciones entre variables si no se modelan explícitamente.

En salud, este modelo se utiliza frecuentemente para estudios de predicción de enfermedades cardiovasculares, diabetes, hipertensión y otros trastornos, a partir de variables como edad, peso, presión arterial, hábitos de vida y antecedentes médicos.

Por otro lado, el árbol de decisión es un modelo que organiza las decisiones en forma de una estructura jerárquica compuesta por nodos. Cada nodo representa una pregunta sobre un atributo del conjunto de datos, y según la respuesta, el árbol dirige hacia una rama u otra hasta llegar a una hoja, que corresponde a una clasificación final.

Su principal fortaleza radica en su facilidad de interpretación. Los árboles de decisión son especialmente útiles en contextos donde se necesita una toma de decisiones comprensible y transparente, como ocurre en la medicina clínica. Son capaces de manejar tanto variables categóricas como numéricas y no requieren una normalización previa de los datos.

No obstante, los árboles de decisión pueden ser inestables y sensibles a pequeñas variaciones en los datos, lo que los hace propensos al sobreajuste si no se emplean técnicas como la poda o la validación cruzada. Por ello, aunque útiles, suelen ser sustituidos por modelos más robustos en aplicaciones más exigentes.

En medicina, se utilizan para representar procesos diagnósticos secuenciales, similares a guías clínicas, facilitando la toma de decisiones médicas en situaciones donde el razonamiento clínico puede estructurarse de forma lógica y ramificada.

En cuanto al método del Random Forest, es un algoritmo de ensamble basado en la construcción de múltiples árboles de decisión, combinando sus resultados para mejorar la precisión del modelo final. Esta técnica se fundamenta en el método de *bagging* (bootstrap aggregating), que consiste en entrenar varios modelos sobre subconjuntos aleatorios del conjunto de datos original y promediar sus predicciones.

Gracias a esta estrategia, el Random Forest reduce significativamente el problema del sobreajuste al que están expuestos los árboles individuales, y mejora la capacidad de generalización del modelo. Además, es eficaz para manejar conjuntos de datos grandes y con múltiples características, incluyendo aquellos con relaciones no lineales y variables irrelevantes.

Una de sus limitaciones es que, al estar compuesto por muchos árboles, el modelo completo puede ser difícil de interpretar. También requiere una mayor cantidad de recursos computacionales y tiempo de entrenamiento en comparación con modelos más simples.

En el ámbito sanitario, Random Forest ha demostrado ser especialmente útil en la predicción de enfermedades complejas como el cáncer, enfermedades neurodegenerativas o trastornos metabólicos, y en la evaluación del riesgo de eventos adversos o reingresos hospitalarios, debido a su alta precisión y capacidad para integrar múltiples fuentes de datos clínicos.

Por último, el clasificador Naive Bayes es un modelo probabilístico basado en el teorema de Bayes, que calcula la probabilidad de que una instancia pertenezca a una clase determinada, dado un conjunto de características. La particularidad de este modelo es que asume que las variables

predictoras son independientes entre sí, lo cual rara vez es cierto en la práctica, pero en muchos casos sigue ofreciendo buenos resultados.

A pesar de esta suposición "naive" (ingenua), este modelo es notablemente eficaz, sobre todo en tareas de clasificación de texto y análisis de sentimientos. Su implementación es sencilla, rápida y eficaz incluso en grandes volúmenes de datos.

En salud, Naive Bayes puede aplicarse, por ejemplo, en la clasificación automática de notas médicas, extracción de información relevante de textos clínicos o predicción de diagnósticos a partir de síntomas codificados. Si bien su rendimiento puede no igualar al de modelos más complejos en ciertos contextos, su bajo costo computacional y facilidad de interpretación lo hacen valioso como modelo base o preliminar.

3.1.2 Modelos de Regresión

El Machine Learning de regresión es una técnica fundamental para la predicción de variables continuas a partir de datos históricos. En el contexto sanitario, este enfoque permite anticipar y cuantificar distintos tipos de resultados clínicos, económicos y operativos. A través de la regresión, se pueden generar estimaciones precisas que facilitan la toma de decisiones tanto clínicas como administrativas.

Una de las aplicaciones más destacadas de la regresión en salud es la predicción de costes asociados a la atención médica. A partir de variables como el diagnóstico, la gravedad del paciente, comorbilidades, duración esperada de hospitalización y tipo de tratamiento, los modelos de regresión pueden predecir el gasto esperado para cada paciente o grupo poblacional.

Esto resulta especialmente útil para la planificación presupuestaria, la optimización del uso de recursos y la evaluación económica de intervenciones clínicas. Por ejemplo, mediante regresión lineal o modelos más avanzados como regresión Lasso, los gestores hospitalarios pueden anticipar los costes de tratamiento de enfermedades crónicas como la diabetes, las enfermedades cardiovasculares o el cáncer.

Además, en sistemas de salud con financiación basada en resultados, estas predicciones permiten ajustar las políticas de reembolso, establecer primas de seguros y evaluar el impacto económico de nuevas terapias.

Otra aplicación clave de la regresión en el entorno hospitalario es la predicción de la demanda asistencial. Los hospitales enfrentan desafíos constantes para gestionar recursos limitados como camas, personal, quirófanos o unidades de cuidados intensivos. Los modelos de regresión permiten anticipar el flujo de pacientes y ajustar la capacidad operativa a la demanda esperada.

Mediante regresión múltiple o modelos no lineales se pueden modelar patrones temporales y estacionales para predecir, por ejemplo, el número de ingresos en urgencias, la ocupación de camas por especialidad, o el volumen de pacientes durante una epidemia.

Estas predicciones son fundamentales para reducir los tiempos de espera, evitar saturaciones del sistema y asignar adecuadamente el personal médico. Además, en contextos de planificación a largo

plazo, ayudan a diseñar estrategias de expansión hospitalaria, distribución territorial de servicios o inversión tecnológica.

La regresión lineal simple es el modelo básico de regresión que estudia la relación entre una única variable independiente y una variable dependiente, bajo la hipótesis de que esta relación es lineal. Este método es útil cuando se busca una estimación simple, como por ejemplo predecir el nivel de un biomarcador en función de la edad. Sin embargo, su simplicidad implica limitaciones, ya que sólo considera un predictor y asume linealidad estricta, lo cual no siempre se cumple en la práctica clínica.

La regresión lineal múltiple es una extensión natural del modelo simple que permite incluir múltiples variables predictoras simultáneamente para explicar la variable objetivo. Este método es ampliamente utilizado en el ámbito sanitario para modelar fenómenos complejos donde influyen varios factores, como la estimación del coste hospitalario en función de edad, sexo, tipo de enfermedad, duración de hospitalización y otros parámetros clínicos. Aunque este método mejora la precisión y permite captar interacciones entre variables, también requiere más datos y un análisis cuidadoso para evitar problemas como la multicolinealidad, que puede distorsionar las estimaciones.

Y, por último, sobre modelos de regresión no lineales el más utilizado es el Random Forest Regressor. Este constituye un algoritmo de aprendizaje conjunto que combina múltiples árboles de decisión para realizar predicciones regresivas y su metodología se basa en agregar predicciones individuales de varios modelos débiles y generar un predictor robusto y preciso.

El algoritmo construye numerosos árboles de decisión utilizando submuestras aleatorias del conjunto de datos original (técnica conocida como bagging o bootstrap aggregating). Cada árbol se entrena con una porción diferente de los datos y, además, en cada nodo del árbol se selecciona aleatoriamente un subconjunto de características para determinar la mejor división. Esta doble aleatorización (en muestras y características) reduce significativamente el sobreajuste y mejora la generalización del modelo. Y para generar la predicción final, Random Forest promedia las predicciones de todos los árboles individuales, lo que resulta en estimaciones más estables y confiables que las obtenidas por un solo árbol de decisión.

En el contexto médico, Random Forest Regressor ofrece múltiples beneficios. aneja eficientemente conjuntos de datos con características heterogéneas, situación común en registros clínicos que combinan variables numéricas (edad, niveles de marcadores clínicos), categóricas (género, tipo de tratamiento) y ordinales (escalas de gravedad). Además, es robusto ante valores atípicos y datos faltantes, problemas frecuentes en bases de datos hospitalarias.

4. Metodología

4.1. Base de datos cáncer de mama

Durante este apartado se profundizará en describir y analizar los procedimientos metodológicos utilizados para abordar los dos objetivos principales: por un lado, el desarrollo de un modelo de clasificación para predecir el diagnóstico de cáncer de mama (tumor benigno o maligno) a partir de variables clínicas; por otro, la elaboración de un modelo de regresión orientado a estimar el coste sanitario asociado a cada paciente en función del estadio de la enfermedad y de los tratamientos recibidos. Para alcanzar estos fines, se han empleado dos bases de datos distintas: una base clínica real proveniente de un hospital africano y un conjunto de datos sintético, elaborado conforme a criterios médicos y epidemiológicos españoles.

Ambas fuentes permiten construir modelos predictivos relejan la necesidad de introducir el Business Intelligence en todas las áreas, en este caso al área médica y de planificación de gastos y recursos mediante su aplicación en la detección temprana y en la optimización de costes en salud pública.

4.1.1. Base de datos de clasificación

Se recolectaron datos de un dataset (fuente: https://data.mendeley.com/datasets/63fpbc9cm4/2) compuesto por 213 observaciones de pacientes, el mismo fue obtenido del registro de cáncer del Hospital Universitario de Calabar durante un período de 24 meses (enero de 2019 a agosto de 2021).

Los datos incluyen 11 características importantes:

- **S/N:** Número identificador del registro. No tiene importancia, se utiliza a modo organizativo.
- **Year:** Año en que se recopiló el dato.
- **Age:** Edad de la paciente (en años).
- **Menopause:** Estado menopáusico de la paciente. Se clasifica en “0” para indicar que no hay menopáusia o “1” cuando si la hay.
- **Tumor Size (cm):** Tamaño del tumor medido en centímetros.
- **Inv-Nodes:** Número de ganglios linfáticos axilares invadidos.
- **Breast:** Indica cuál de los pechos es afectado por el cáncer. Se clasifica en “Left” para el izquierdo o “Right” para el derecho.
- **Metastasis:** Presencia de metástasis. Se clasifica en “0” para indicar que no hay metástasis o “1” cuando si la hay.
- **Breast Quadrant:** Cuadrante del seno donde se localiza el tumor. Se puede clasificar en “Upper outer” (superior externo), “Upper inner” (superior interno), “Lower outer” (inferior externo), “Lower inner” (inferior interno).
- **History:** Indica si el paciente incluye historial de enfermedades mamarias. Se clasifica en “0” para indicar que no hay historial o “1” cuando lo hay.

- **Diagnosis Result:** Variable resultante para saber si el paciente tiene o no un tumor benigno o maligno. Se clasifica en “Benign” (Tumor benigno) o “Malignant” (Tumor maligno, tiene cáncer).
- **Smoke:** Indica si el paciente es fumador. Se clasifica en “0” para indicar que no fuma o “1” para los pacientes que si suelen fumar.
- **Sugar:** Indica si el paciente es diabético. Se clasifica en “0” para indicar que no lo es o “1” cuando si es diabético.

4.1.2. Base de datos de modelo de regresión

Para la realización del modelo de regresión no ha sido posible obtener bases de datos privadas de los centros médicos por lo que se ha elaborado un dataset propio, por lo que es importante detallar que exclusivamente ha sido diseñado para fines académicos. La información que dispone ha sido simulada teniendo en cuenta y según los criterios médicos realistas basados en la literatura científica, guías de estadificación TNM, y datos epidemiológicos nacionales e internacionales, de forma que su estructura y distribución sean representativas de la situación española actual del cáncer de mama.

Se persiguen dos objetivos claros para este conjunto de datos, en primer lugar, ofrecer una base sólida y realista para la elaboración de un modelo predictivo de gastos público y con la clara intención de conseguir una disminución del gasto gracias a la detección temprana. En segundo lugar, la creación de unos Dashboard que servirá de panel de control para entender cómo se encuentra la enfermedad y el gasto que conlleva, con el fin de poder tomar decisiones basadas en datos de forma más ágil.

La elaboración de este apartado sigue dos áreas relevantes: los costes y los pacientes.

4.1.2.1 Base de datos de Costes

Como parte fundamental para obtener conclusiones efectivas, se tuvo que recolectar información de los costes relacionados al cáncer de mama.

Para el mismo se creó un informe, compaginando distintas fuentes, recolectando la información de los últimos años para entender cómo están distribuido esos datos. El mismo contiene un análisis minucioso del análisis global de los costes en España sobre esta enfermedad, su impacto económico, detalles de los costes directos e indirectos, costes por estadío del paciente, incluyendo para cada uno los tipos de servicios y/o tratamientos (diagnósticos, radioterapia, quimioterapia, cirugías, mamografías, biopsias, etc.) y los valores del mercado.

También se analizaron los valores de las consultas médicas (primeras consultas, consultas generales y las de urgencia), costes por hospitalización en distintas localidades de España.

Una vez recolectada toda esa información se realizó la construcción de un dataset que permita leer los datos de una forma cómoda para luego realizar todas las tareas que requieren el análisis estadístico de una regresión. El dataset se distribuyó en 4 variables:

- **Categoría de Coste:** Cada coste se clasifica por categoría la cual permite encontrar fácilmente los costos correspondientes para cada caso de cada paciente, por ejemplo: costes relacionados con el estadío, con quimioterapia, radioterapia, etc.
- **Descripción:** Explica el detalle del coste.
- **Coste Estimado:** Es el coste medio de todos los datos obtenidos en el informe realizado (distintas clínicas, públicas, privadas, etc.).
- **Frecuencia / Unidad de Coste:** Este dato es importante para entender si el coste corresponde a un coste fijo de por vida, o a un coste mensual, diario, por consulta, entre otros.

4.1.2.1 Base de datos de Pacientes

En esta ocasión cada fila representa un paciente individual, y cada columna recoge una variable relevante para el diagnóstico, evolución, tratamiento o seguimiento de los costes directos del cáncer de mama. Dejamos fuera del análisis los costes indirectos o costes invisibles que se comentaron en el apartado 2.2.

- **ID:** Identificador único de cada paciente en el conjunto de datos (P01, P02, …).
- **Sexo:** Género del paciente (en este caso, todas son "Mujer").
- **Edad:** Años cumplidos del paciente en la fecha actual (15 jun 2025).
- **Ciudad:** Ciudad que atiende al paciente de forma habitual, donde esta asignado el caso.
- **Fecha detección:** Día en que se diagnosticó por primera vez el cáncer de mama.
- **Años con la enfermedad:** Tiempo transcurrido, en años y un decimal aproximado, desde la fecha de detección hasta el 15 jun 2025.
- **Tipo de cáncer:** Nombre completo del subtipo de cáncer de mama diagnosticado (sin siglas).
- **Estadio:** Grado de avance del cáncer (0, I, II, III o IV) según la clasificación clínica TNM.
- **Lógica de asignación:** Explicación en lenguaje sencillo de por qué al caso se le asignó ese estadio concreto, sin terminología técnica.
- **Tumor (cm):** Tamaño máximo medido del tumor primario en centímetros.
- **Ganglios axilares:** Número de ganglios linfáticos de la axila que mostraron células cancerosas.
- **Metástasis:** Si el cáncer se ha extendido a órganos lejanos al pecho ("Sí" / "No").
- **Invasión pared/piel:** Si el tumor ha infiltrado la piel o la pared torácica ("Sí" / "No").
- **Nºmamografías realizadas:** Número total de mamografías administradas.
- **Visitas médicas:** Número de consultas presenciales u online con profesionales de salud durante el primer año de tratamiento o el último año controlado.
- **Coste unitario consulta (€):** coste/ cantidad en euros del valor de una consulta médica
- **Coste consultas (€):** total en euros de las consultas médicas que ha gastado el paciente.
- **Biopsias:** Cantidad de procedimientos de extracción de tejido realizados para diagnóstico o control.
- **Coste unitario biopsia (€):** coste / cantidad en euros del valor de una biopsia.
- **Coste biopsias (€):** Total en euros que ha gastado el paciente de las biopsias realizadas.
- **Radio:** Indica si el paciente recibió tratamiento de radioterapia ("Sí" / "No").
- **Nº sesiones RT:** Número total de fracciones (sesiones) de radioterapia administradas.

- **Coste unitario radioterapia (€):** coste en euros del valor de una radioterapia.
- **Coste radioterapia (€):** total en euros que ha gastado el paciente de las sesiones realizadas en radioterapia.
- **Quimio:** Indica si el paciente recibió tratamiento de quimioterapia (“Sí” / “No”).
- **Nº ciclos QT (llevados):** Cantidad de ciclos de quimioterapia completados; si Continúa en quimio = Sí, refleja los ciclos ya realizados hasta la fecha.
- **Continúa en quimio:** Muestra si la quimioterapia sigue activa hoy en día (“Sí”) o ya terminó (“No”).
- **Coste unitario quimioterapia (€):** coste en euros del valor de una quimioterapia.
- **Coste quimioterapia (€):** total en euros que ha gastado el paciente de las sesiones realizadas en quimioterapia.
- **Extracción:** Indica si se practicó cirugía para extraer el tumor o la mama (“Sí” / “No”).
- **Tipo extracción:** Tipo de cirugía realizada: Tumorectomía (solo el tumor), Mastectomía (toda la mama) o No hubo.
- **Coste unitario extracción (€):** Monto en euros del valor de una extracción.
- **Coste extracción (€):** Monto en euros que ha gastado el paciente en extracciones.
- **Reconstrucción mamaria:** Si hubo reconstrucción mamaria tras mastectomía o no.
- **Coste total paciente (€):** Monto en euros del total gastado por el paciente.

En todo momento se garantiza la coherencia entre variables mediante reglas clínicas derivadas de guías como el sistema TNM (Clasificación de tumores malignos) o los informes que ofrece la asociación del cancer de mama, por ello:

- El estadio se asigna considerando tamaño tumoral, afectación ganglionar y presencia de metástasis.
- Los estadios también se distribuyen de forma realista:
    - Estadio 0: 15%
    - Estadio I: 45%
    - Estadio II: 28%
    - Estadio III: 7%
    - Estadio IV: 5%
- El uso de quimioterapia o radioterapia depende del estadio, subtipo histológico y edad.
- La reconstrucción mamaria solo se considera en pacientes mastectomizadas no metastásicas.
- La distribución por tipo de cáncer respeta las frecuencias estimadas reales:
    - Carcinoma ductal invasivo: ~70%
    - Carcinoma lobulillar invasivo: ~10%
    - Cáncer in situ: ~15%
    - Tipos raros (Paget, inflamatorio, filoides, angiosarcoma): <5% en conjunto

Toda esta información ha sido mencionada durante el apartado 2 del presente TFM.

5. Análisis de las Bases de datos

5.1 Análisis preliminar

El propósito de este apartado es presentar los resultados del Análisis Exploratorio de Datos (EDA) realizado sobre tanto conjunto de datos de cáncer de mama como los datos sobre los pacientes y sus gastos hospitalarios. El EDA es un paso crucial que permite comprender las características principales del dataset, identificar la necesidad de limpieza, descubrir patrones y visualizar las relaciones entre las distintas variables antes de aplicar modelos de clasificación o modelos de regresión.

5.1.1 Exploración de datos para la clasificación

En este apartado se desarrolla un análisis exploratorio exhaustivo para preparar los datos e identificar patrones que faciliten la construcción de un modelo de clasificación binaria entre diagnósticos benignos y malignos. Los objetivos específicos abarcan la evaluación de la calidad y completitud de los datos, la identificación y tratamiento de valores faltantes e inconsistencias, el análisis de la distribución de variables individuales y sus relaciones, la determinación de las variables más predictivas mediante análisis de correlación, y la aplicación de técnicas de reducción dimensional para validar la separabilidad de clases. A continuación, se detallan cada uno de estos objetivos por separado.

El dataset utilizado proviene del archivo breast-cancer-dataset-final.csv, descrito en el apartado 4.1.1 de este documento, que contiene datos clínicos de pacientes con cáncer de mama. La variable objetivo del análisis es *Diagnosis Result*, que clasifica los casos como benignos o malignos, estableciendo así un problema de clasificación binaria supervisada. Este fichero presenta características estructurales bien definidas, con 213 registros de pacientes distribuidos a través de 13 características clínicas y demográficas. Esta configuración establece un problema de clasificación binaria supervisada con una densidad de información adecuada para el análisis exploratorio y el modelado posterior.

En cuanto a las variables, el csv contiene una combinación diversa de tipos de variables que capturan diferentes aspectos de la condición clínica. Las variables cuantitativas continuas incluyen la edad de la paciente medida en años y el tamaño tumoral expresado en centímetros, ambas fundamentales para la caracterización clínica.

Las variables cuantitativas discretas abarcan el estado menopáusico, el hábito tabáquico, el nivel de glucosa, los nódulos linfáticos invadidos, la presencia de metástasis y el historial familiar, todas codificadas binariamente para facilitar el análisis cuantitativo.

Las variables categóricas incluyen la lateralidad mamaria que distingue entre mama izquierda y derecha, el cuadrante mamario que especifica la localización anatómica en cuatro categorías, y la variable objetivo que clasifica el diagnóstico como benigno o maligno.

Una vez descrito a grandes rasgos el dataset se ha realizado un análisis estadístico inicial el cual revela características importantes de la población estudiada. La edad promedio de las pacientes es de 39.78 años con una desviación estándar de 14.10 años, indicando una distribución relativamente amplia que abarca desde pacientes jóvenes de 13 años hasta pacientes de 77 años.

El estado menopáusico muestra que aproximadamente el 67% de las pacientes se encuentran en estado post-menopáusico, lo cual es consistente con el perfil epidemiológico del cáncer de mama. El hábito tabáquico está presente en el 35% de las pacientes, mientras que solo el 13% presenta niveles elevados de glucosa, proporcionando información sobre factores de riesgo adicionales.

Métrica	Age	Menopause	Smoke	Sugar
count	213,0	213,0	213,0	213,0
mean	39,78	0,67	0,35	0,13
std	14,10	0,47	0,48	0,33
min	13,0	0,0	0,0	0,0
25%	30,0	0,0	0,0	0,0
50%	40,0	1,0	0,0	0,0
75%	49,0	1,0	1,0	0,0
max	77,0	1,0	1,0	1,0

Tabla 21 Resumen estadístico de las variables de Clasificación

Fuente: Elaboración propia con Python

Durante la inspección inicial se identificó una problemática significativa relacionada con la codificación de valores faltantes, que aparecían representados por el carácter especial '#' en lugar de utilizar la notación estándar de valores nulos. Esta inconsistencia requirió un tratamiento específico para asegurar la integridad del análisis posterior.

La columna *Breast* presentaba seis valores faltantes codificados de esta manera, y se detectaron inconsistencias adicionales en la representación de datos categóricos que podrían afectar la calidad del análisis. La estrategia de limpieza implementada comenzó con la conversión sistemática de todos los marcadores '#' a valores NaN para permitir su tratamiento mediante métodos estándar de manejo de datos faltantes.

Posteriormente se procedió a la eliminación de registros incompletos mediante la función dropna(), asegurando que el dataset final contuviera únicamente observaciones completas. Esta aproximación, aunque reduce el tamaño muestral, garantiza la integridad de los análisis posteriores y evita sesgos introducidos por métodos de imputación inapropiados.

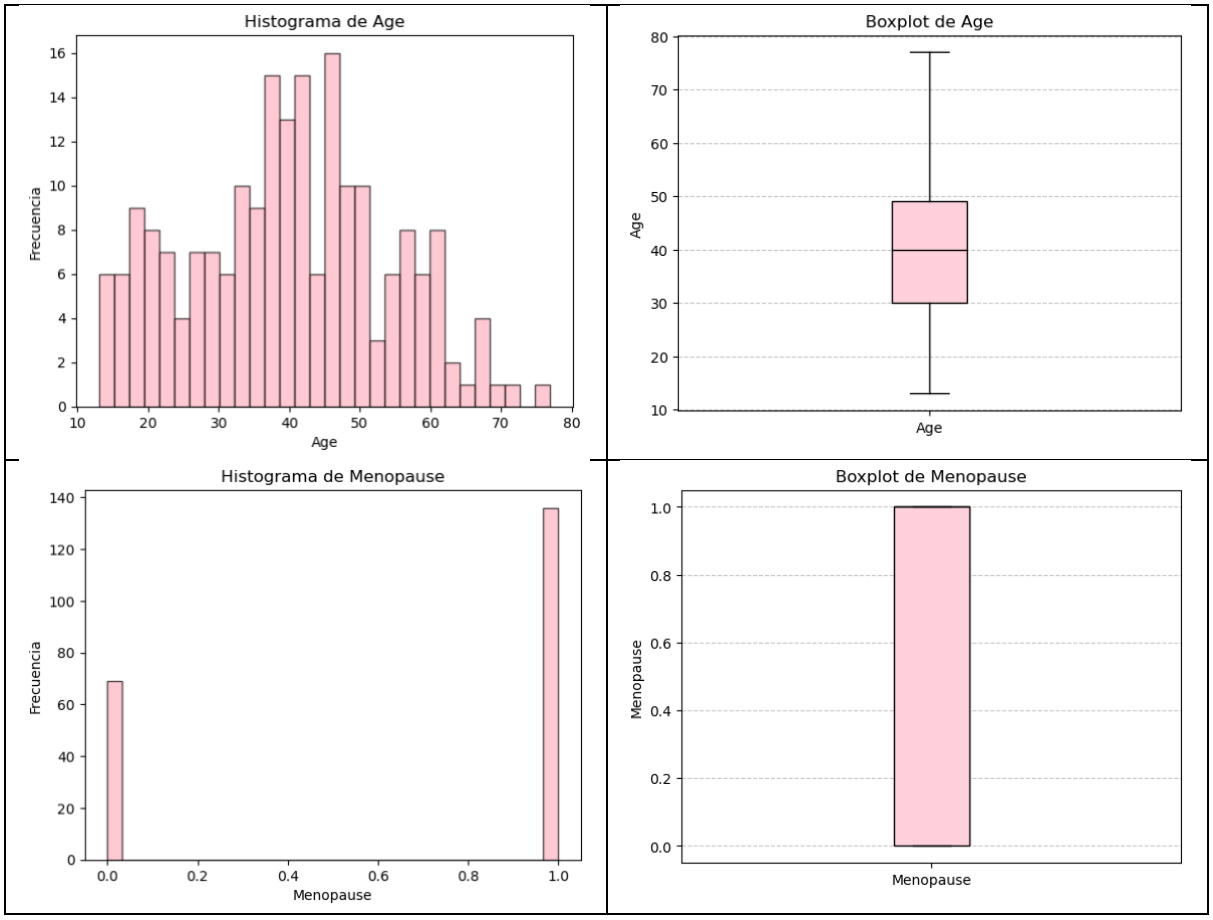
Siguiendo con la parte del tratamiento y limpieza del dataset, se procede a la transformación de variables. Varias columnas que contenían información numérica estaban incorrectamente tipificadas como objetos, requiriendo conversión explícita a tipos numéricos apropiados. Las variables *Age*, *Tumor Size*, e *Inv-Nodes* fueron convertidas a tipos numéricos, mientras que *Metastasis* e *History* fueron estandarizadas como variables binarias.

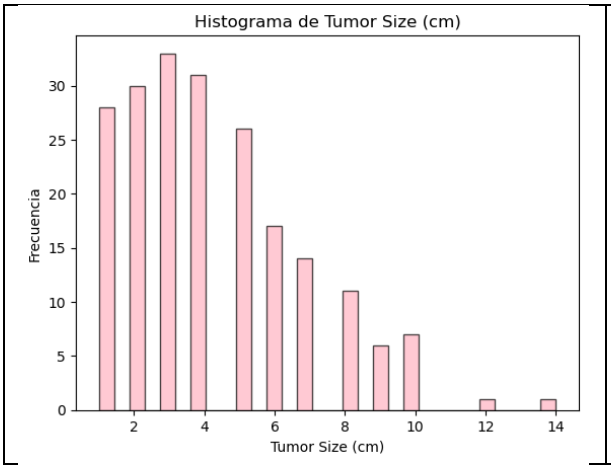
La codificación de variables categóricas requirió tratamiento especializado para cada caso. La variable *Breas*t, que indica lateralidad, fue transformada mediante codificación one-hot, creando la variable *Breast_Right* donde el valor 1 representa mama derecha y 0 representa mama izquierda. Esta transformación preserva la información categórica mientras permite su inclusión en análisis cuantitativos.

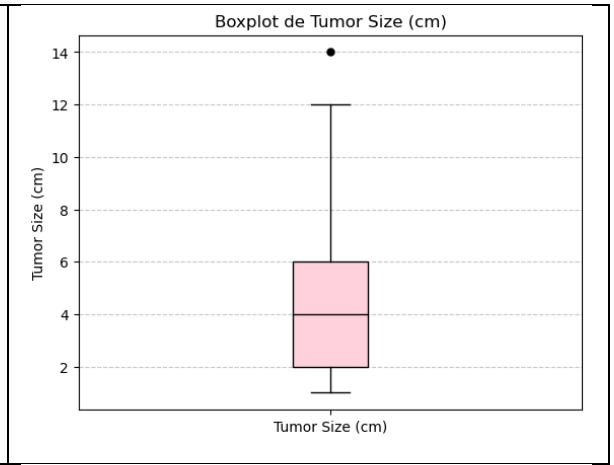
La variable *Breast Quadrant* presentó inconsistencias en las etiquetas, con variaciones como *Upper outer* y *Upper outer* que requerían estandarización. Después de la limpieza, se implementó un mapeo numérico donde *Upper outer* corresponde a 1, *Lower outer* a 2, *Upper inner* a 3, y *Lower inner* a 4.

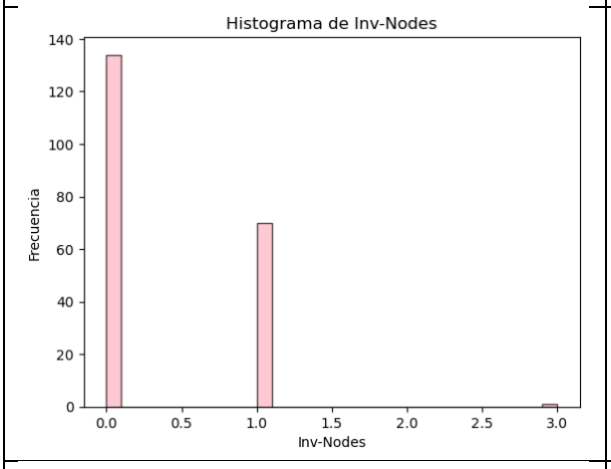
La variable objetivo *Diagnosis Result* fue codificada binariamente asignando el valor 1 a casos malignos y 0 a casos benignos, estableciendo así la convención estándar para problemas de clasificación binaria donde la clase positiva representa el evento de interés.

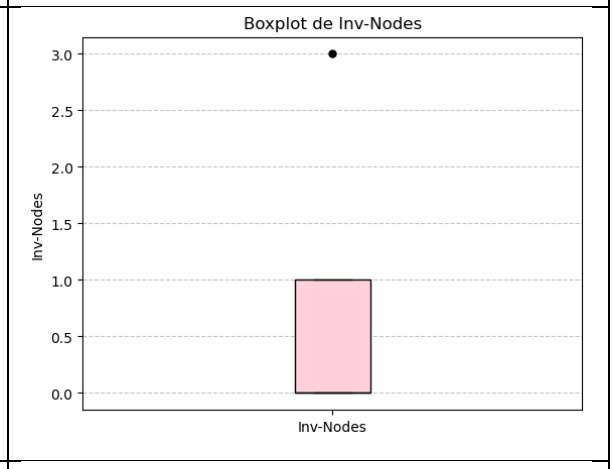
Para finalizar con el parte del proceso de limpieza, se decidió eliminar las variables *S/N* y *Year* por considerarse irrelevantes para el análisis de clasificación dado que la primera actúa como identificador sin valor predictivo, mientras que la segunda representa información temporal que no contribuye al modelo de clasificación que se pretende construir.

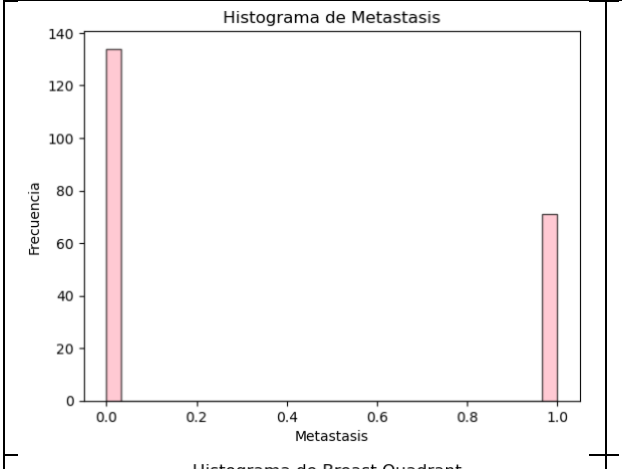


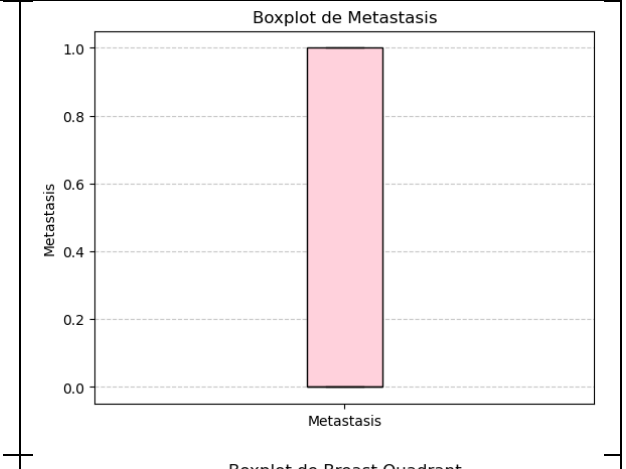

Histograma de Tumor Size (cm)
Frecuencia
Tumor Size (cm)

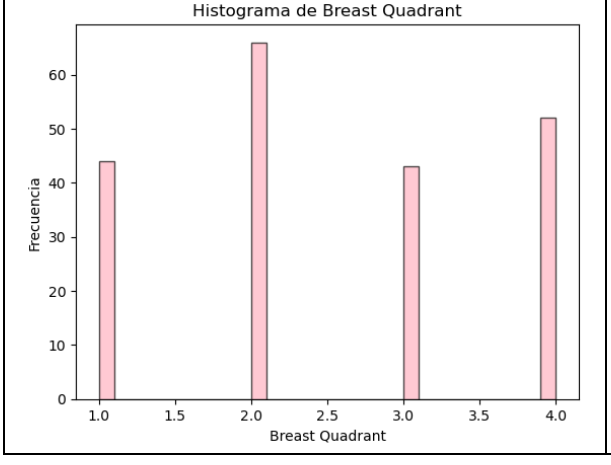

Boxplot de Tumor Size (cm)
Tumor Size (cm)
Tumor Size (cm)

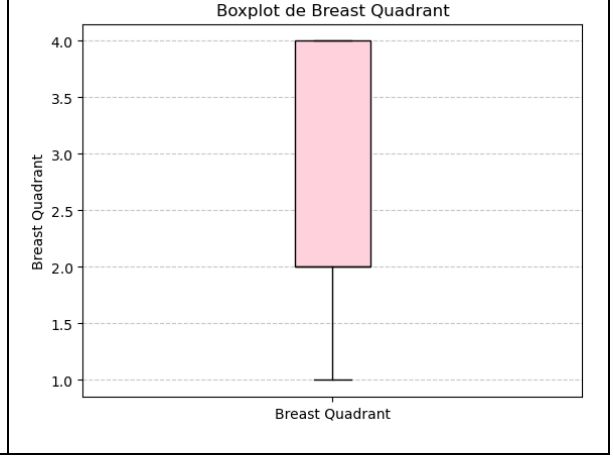

Histograma de Inv-Nodes
Frecuencia
Inv-Nodes


Boxplot de Inv-Nodes
Inv-Nodes
Inv-Nodes


Histograma de Metastasis
Frecuencia
Metastasis


Boxplot de Metastasis
Metastasis
Metastasis


Histograma de Breast Quadrant
Frecuencia
Breast Quadrant


Boxplot de Breast Quadrant
Breast Quadrant
Breast Quadrant

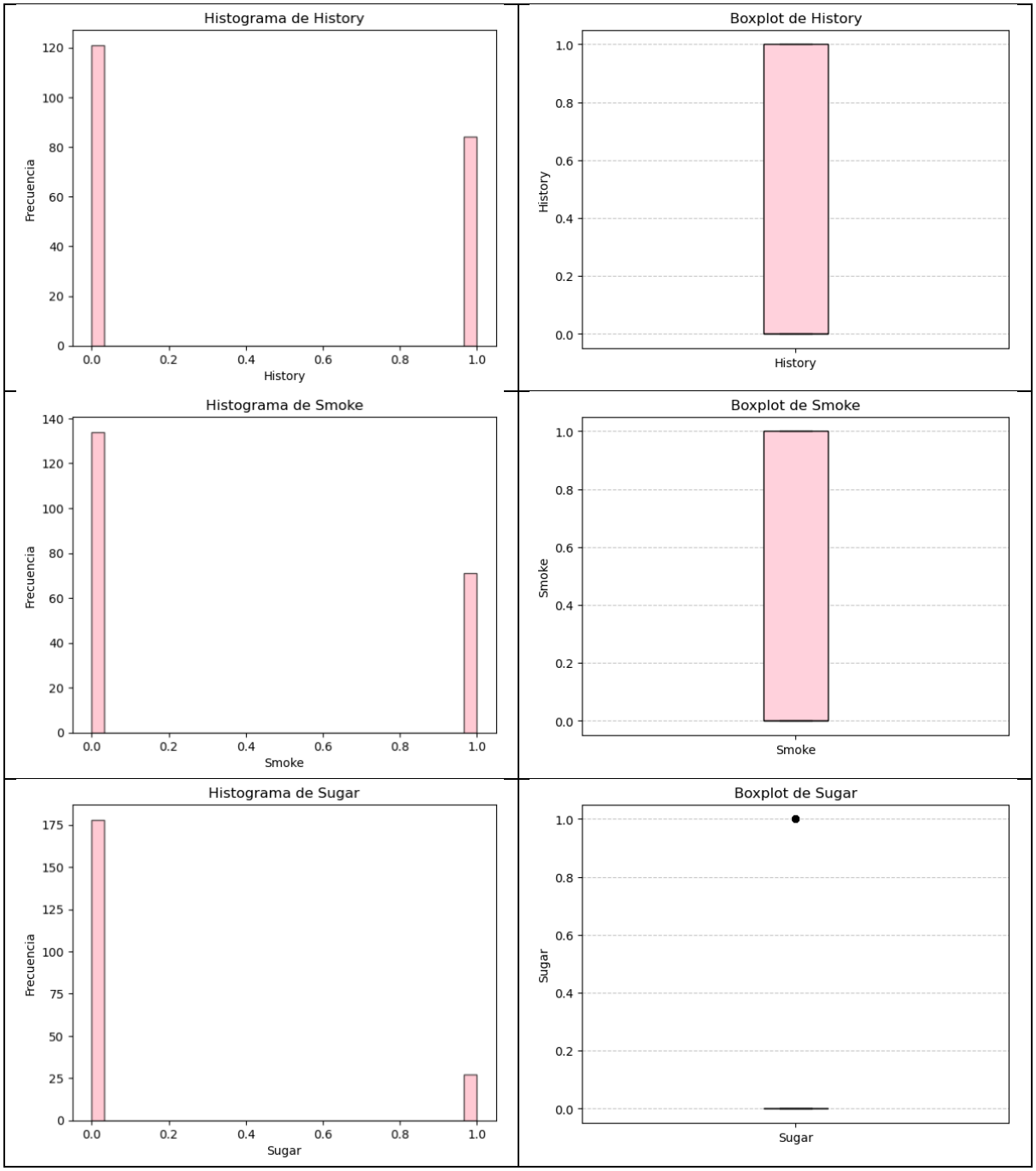


Tabla 22 Gráficos exploratorios del modelo de Clasificación

Fuente: Elaboración propia con Python

Tras analizar los histogramas y los boxplots realizados se observa que la distribución de la variable Age revela una distribución aproximadamente normal con ligera asimetría positiva, concentrándose la mayor densidad de observaciones en el rango de 30 a 50 años. Esta distribución es consistente con la epidemiología conocida del cáncer de mama, donde la incidencia aumenta con la edad, pero mantiene una presencia significativa en grupos de edad media.

El rango completo de edades, desde 13 hasta 77 años, sugiere que el dataset incluye casos tanto de cáncer de mama juvenil como de presentación en edades avanzadas, proporcionando una representación amplia del espectro etario de la enfermedad.

La variable Tumor Size presenta una distribución marcadamente asimétrica hacia la derecha, con una concentración significativa de casos en tamaños pequeños menores a 3 centímetros. Esta característica indica que la mayoría de los casos fueron detectados en estadios tempranos, lo cual es favorable desde una perspectiva clínica.

La presencia de algunos casos con tumores de gran tamaño actúa como valores atípicos en la distribución, representando casos de detección tardía o tumores de crecimiento agresivo. Esta variabilidad en el tamaño tumoral proporciona un rango dinámico importante para la discriminación entre casos benignos y malignos.

La distribución de diagnósticos en el dataset muestra la proporción relativa entre casos benignos y malignos, información importante para evaluar el balance de clases y determinar si existen sesgos que podrían afectar el rendimiento de modelos de clasificación. La distribución entre mama derecha e izquierda puede proporcionar información sobre posibles asimetrías en la presentación de la enfermedad, mientras que la localización por cuadrantes puede revelar patrones relacionados con el espacio que influyan en el pronóstico o las características del tumor.

A continuación, se realizan gráficos comparativos entre la edad y el tamaño del tumor frente a la variable objetivo, el resultado del diagnóstico, para estudiar la relación que existe entre ellas.

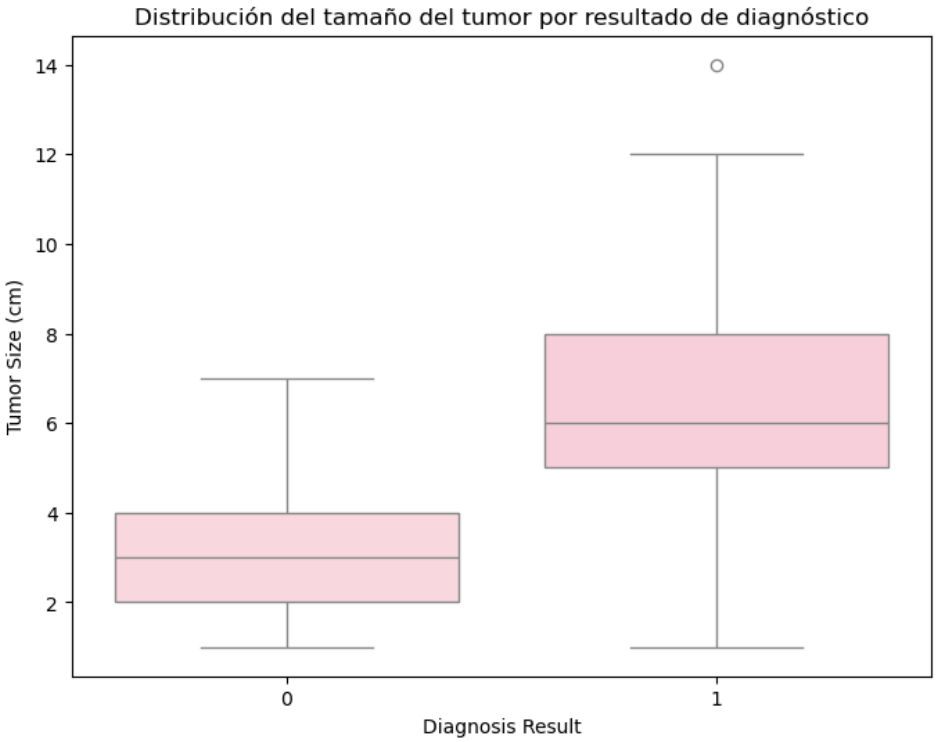


Tabla 23 BoxPlot del tamaño del tumor y su diagnóstico

Fuente: Elaboración propia con Python.

El análisis de la relación entre el tamaño tumoral y el resultado diagnóstico revela un patrón claro donde los tumores malignos presentan significativamente mayor tamaño promedio comparado con los benignos. Sin embargo, existe una zona de solapamiento que indica que el tamaño tumoral por sí

solo no es completamente discriminativo lo cual sugiere la necesidad de considerar variables adicionales para una clasificación óptima.

Esta relación desde el punto de vista de la medicina es coherente, ya que los tumores malignos tienden a crecer más agresivamente y alcanzar mayores tamaños antes de la detección. No obstante, la presencia de outliers a nivel estadístico en el dataset y que traducidos a información clínica muestran la existencia de tumores malignos pequeños y tumores benignos grandes subraya la importancia de un enfoque multivariado para la clasificación.

La relación entre edad y diagnóstico muestra una tendencia donde los diagnósticos malignos están asociados con edades ligeramente superiores. Este patrón refleja el incremento del riesgo de que la tipología del tumor sea maligna con la edad. La correlación, aunque presente, es moderada, indicando que la edad es un factor de riesgo importante, pero, al igual que se mencionaba con el tamaño del tumor, no decisivo.

En un análisis exploratorio de datos, la matriz de correlaciones facilita el identificar qué relaciones existen entre las variables y ayuda a descubrir multicolinealidad, permitiendo decidir si eliminar variables redundantes o transformarlas.

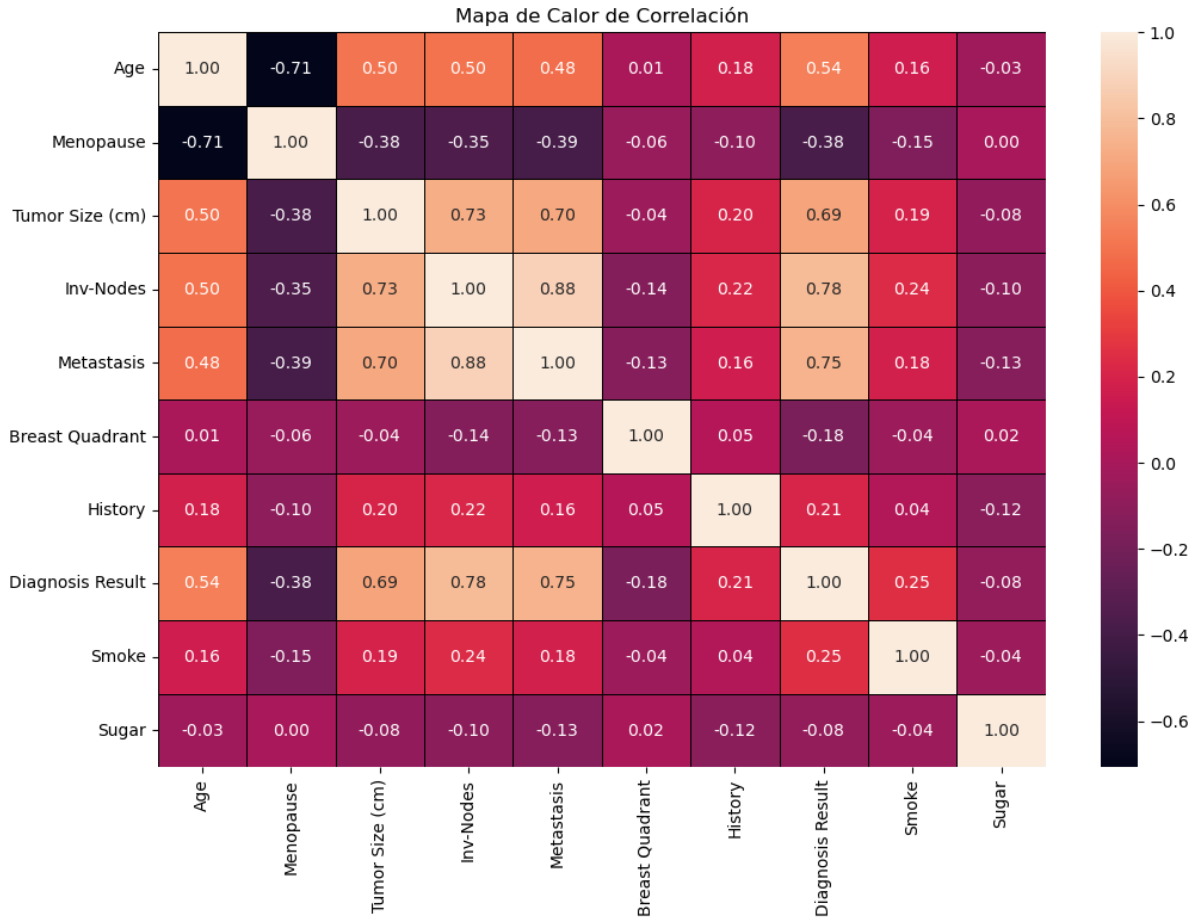


Tabla 24 Mapa de calor del modelo de Clasificación

Fuente: Elaboración propia con Python

Como bien se puede observar, el mapa de calor apoya que las variables con mayor correlación con el resultado diagnóstico son precisamente aquellas identificadas en el análisis anterior. El tamaño tumoral muestra la correlación positiva más fuerte con el diagnóstico maligno, seguido por la edad con una correlación positiva moderada.

Las variables relacionadas con la extensión de la enfermedad, como la presencia de metástasis y la invasión de nódulos linfáticos, también presentan correlaciones positivas esperadas con el diagnóstico maligno lo cual tiene sentido desde una perspectiva clínica, ya que tanto la metástasis como la invasión nodal son características definitorias del tumor maligno.

Por tal de concluir esta fase de exploración de los datos se ha realizado un análisis de componentes principales (PCA). Esta técnica pretende de reducir la dimensionalidad transformando las variables originales en combinaciones lineales ortogonales llamadas componentes principales. Cada componente captura la mayor varianza posible bajo la restricción de ser independiente de los anteriores, priorizando la información más relevante. Para realizarlo se ha desarrollado el siguiente código python:

Los resultados del PCA revelan que el primer componente principal explica el 96.6% de la varianza total en los datos, mientras que el segundo componente contribuye con un 2.4% adicional. Esta concentración extrema de varianza en los primeros dos componentes indica que la estructura de los datos es relativamente simple y que la mayoría de la información discriminativa puede ser capturada en un espacio bidimensional.

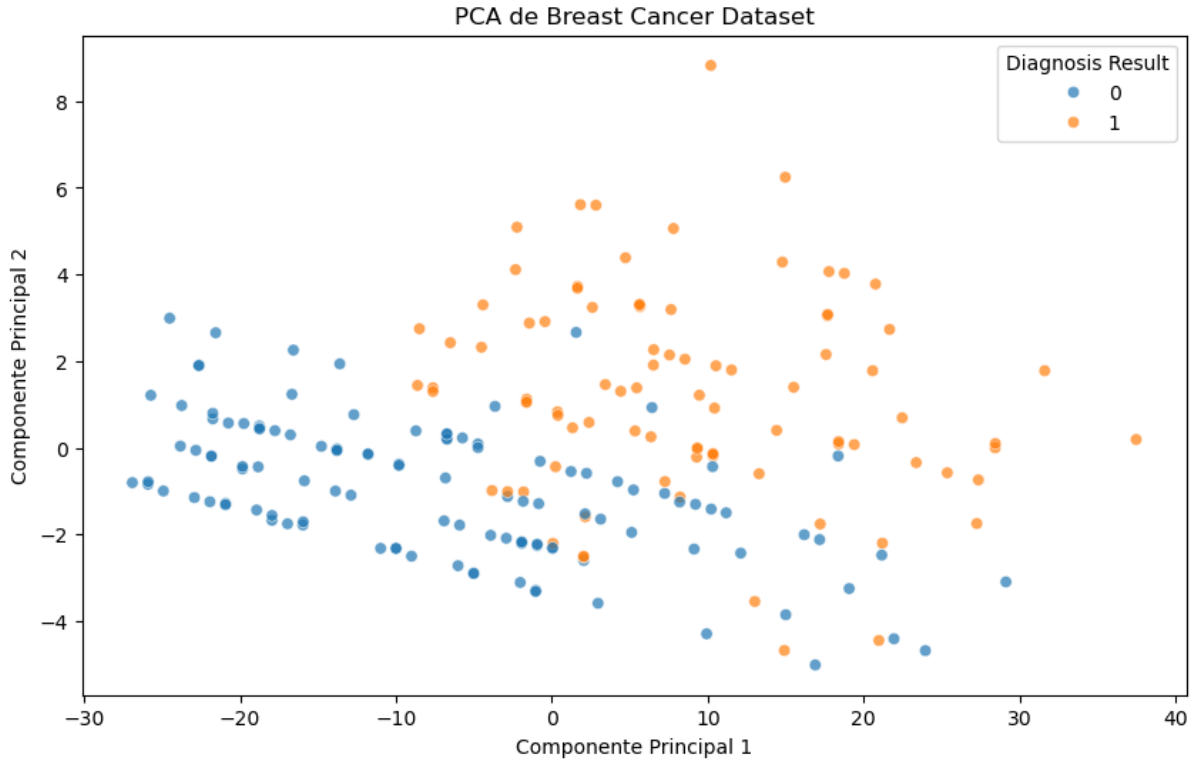


Tabla 25 Distribución del diagnóstico según el PCA

Fuente: Elaboración propia con Python

La interpretación de los componentes indica que el primer componente está dominado por la contribución de la variable edad, representando el eje principal de variabilidad demográfica en la población estudiada. El segundo componente viene influenciado por el tamaño tumoral, capturando la variabilidad relacionada con las características morfológicas del tumor.

Tras estas primeras con conclusiones del análisis se puede observar cómo los resultados del PCA se encuentran alineados con los análisis bivariantes vistos anteriormente.

Además, se ha considerado interesante realizar un círculo de correlación ya que proporciona una representación visual clara de cómo cada variable original contribuye a los componentes principales del PCA. La edad aparece como el vector de mayor magnitud en el primer componente, confirmando su papel dominante en la variabilidad principal de los datos.

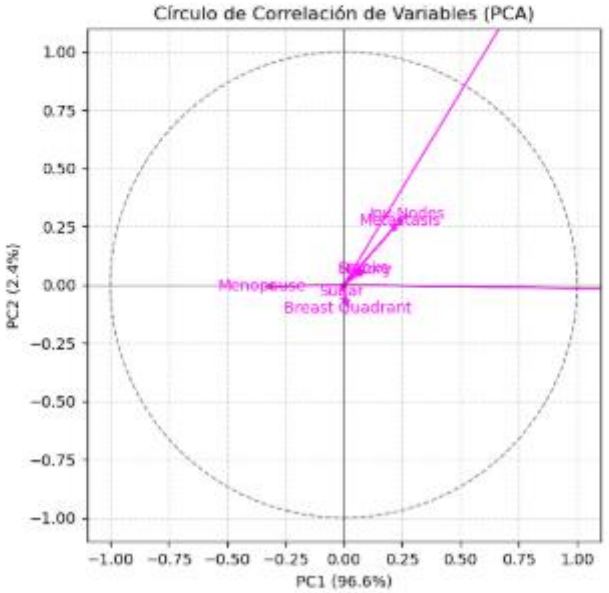


Tabla 26 Implicación de las variables según el PCA

Fuente: Elaboración propia con Python

El tamaño tumoral se presenta como un componente perpendicular significativo, indicando que su contribución es ortogonal a la de la edad. Esta ortogonalidad confirma que ambas variables aportan información independiente y complementaria para la caracterización de los casos.

5.1.2 Exploración de datos para la regresión

Del mismo modo que para la base de datos utilizados para construir el modelo de clasificación se ha realizado un análisis exploratorio de los datos, a continuación, se procede a realizar el mismo estudio para la información recogida en el dataset que posteriormente se utilizará para entrenar modelos de regresión.

El dataset dataset_cancer_mama_300_ciudad_columnas_ordenadas.csv. como bien se ha explciado en el apartado 4.1.2.1 Pacientes, recoge información de 300 pacientes con cáncer de mama, combinando variables clínicas, demográficas y económicas. Se compone de 19 variables numéricas y 16 variables categóricas. Las variables numéricas incluyen edad, número de ganglios axilares afectados, visitas médicas, número de biopsias, sesiones de radioterapia y ciclos de quimioterapia, así como los distintos costes en euros por cada procedimiento médico (consultas, biopsias, radioterapia, quimioterapia, extracción, reconstrucción) y el coste total del tratamiento por paciente, que es la variable objetivo para predicción. Por otro lado, las variables categóricas comprenden datos como sexo, ciudad, tipo de cáncer, estadio, presencia de metástasis, tipo de extracción y si continúa en tratamiento.

En este caso no ha sido necesaria la limpieza de datos puesto que los datos han sido creados por los autores de este trabajo a través de prompts a la Inteligencia Artificial detallados sobre qué información plasmar. Es por eso por lo que directamente se ha realizado un análisis estadístico inicial el cual facilitará la comprensión de las características de los pacientes con cáncer de mama.

Métrica	Edad	Años con la enfermedad	Estadio	Tumor (cm)	Ganglios axilares	Nº mamografías realizadas	Visitas médicas	Coste unitario consulta	Coste consultas	Biopsias
count	300	300	300	300	300	300	300	300	300	300
mean	54,65	2,17	1,42	2,45	1,71	3,35	22,90	49,00	1122,26	1,37
std	21,32	1,27	0,99	2,04	3,28	1,19	7,48	0,00	366,49	0,48
min	18	0,4	0	0,1	0	1	8	49	392	1
25%	35,75	1,3	1	0,9	0	3	18	49	882	1
50%	56	1,9	1	1,6	0	3	24	49	1176	1
75%	73	2,7	2	3,63	2	4	29	49	1421	2
max	89	7,6	4	7,9	15	8	35	49	1715	2

Tabla 27 Exploración del modelo de regresión

Fuente: Elaboración propia con Python

Métrica	Coste unitario biopsia	Coste biopsias	Nº sesiones RT	Coste unitario radioterapia	Coste radioterapia	Nº ciclos QT (llevados)	Coste unitario quimioterapia	Coste quimioterapia	Coste unitario extracción	Coste extracción	Coste total paciente
count	300	300	300	300	300	300	300	300	300	300	300
mean	230	314,33	24,35	280	6817,07	3,80	8000	30373,33	3760	3640	42267,00
std	0	111,02	6,73	0	1884,08	4,02	0	32182,08	972,39	1214,63	32421,94
min	230	230	0	280	0	0	8000	0	3000	0	7969
25%	230	230	22	280	6160	0	8000	0	3000	3000	11073,25
50%	230	230	26	280	7280	4	8000	32000	3000	3000	44903,5
75%	230	460	28	280	7840	7	8000	56000	5000	5000	69922,25
max	230	460	35	280	9800	20	8000	160000	5000	5000	161455

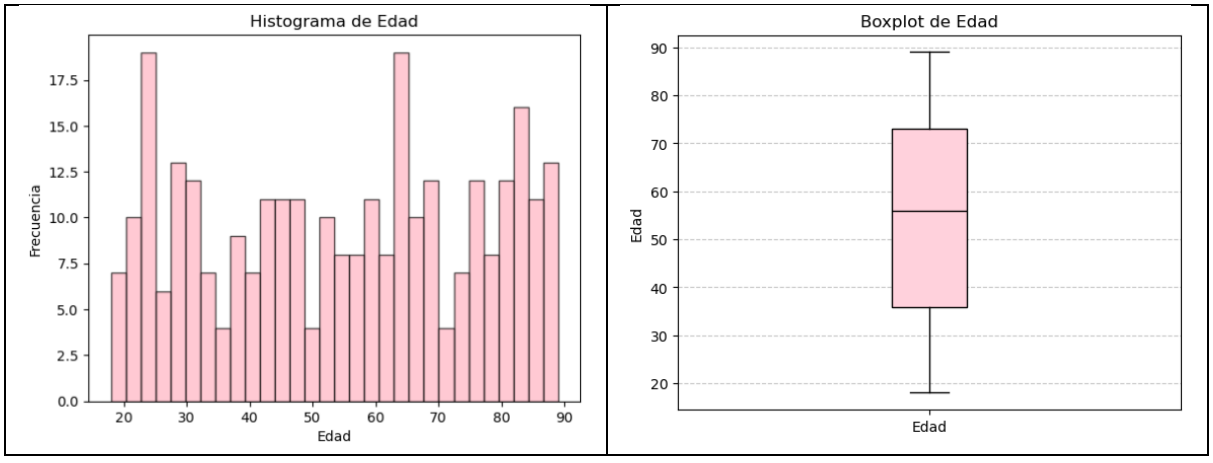
Tabla 28 Exploración 2 del modelo de regresión

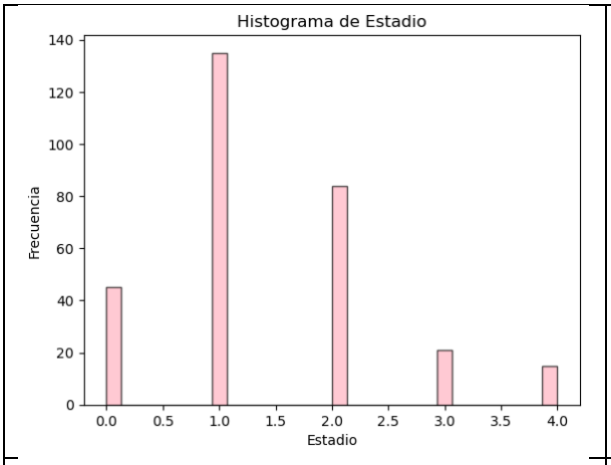
Fuente: Elaboración propia con Python

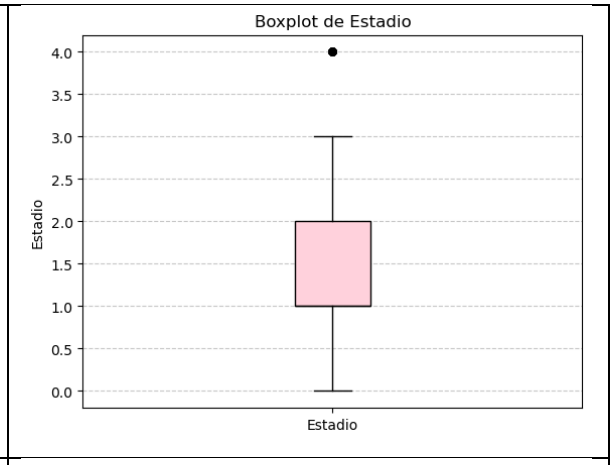
Como se puede observar en las siguientes tablas, la muestra estudiada presenta un perfil demográfico diverso con una edad promedio de 48,5 años y una desviación estándar de 21,32 años, abarcando desde pacientes jóvenes de 18 años hasta adultos mayores de 89 años. Esta variabilidad de edad se complementa con un tiempo promedio de enfermedad de 2,17 años, indicando una mezcla de pacientes desde diagnósticos recientes hasta casos con 7,6 años de evolución.

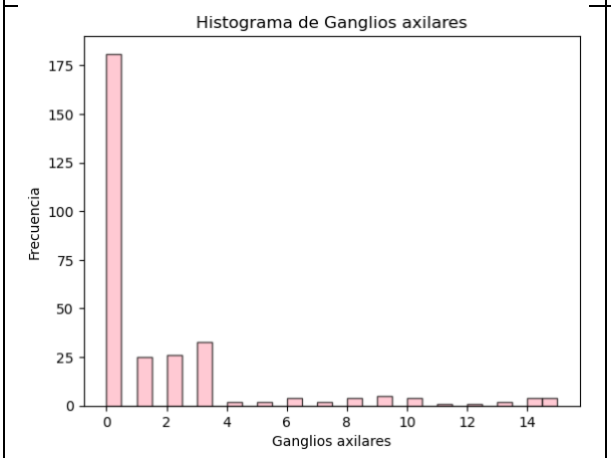
Desde la perspectiva de los datos clínicos, el estadio del tumor presenta una media de 1,42 centímetros con una desviación estándar baja (0,99), donde el 75% de los pacientes se encuentra en estadios ≤2 centímetros. Esta concentración en estadios tempranos se correlaciona con la existencia de ganglios axilares afectados, observándose una media de 1,71 de estos y el 75% de los pacientes con 2 o menos ganglios comprometidos.

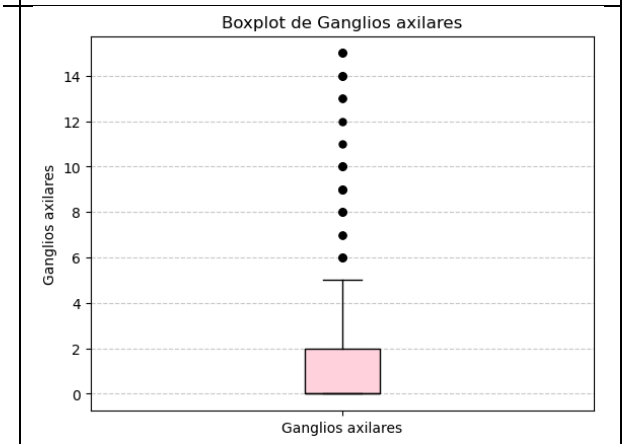
El número de visitas médicas presenta una alta variabilidad, correlacionándose con costes de consulta igualmente diversos. Esta dispersión sugiere diferentes intensidades de seguimiento y protocolos de atención según la complejidad del caso, estadio de la enfermedad y respuesta al tratamiento. La variabilidad extrema en los costos de tratamiento, especialmente evidente en quimioterapia donde la cantidad oscila desde la ausencia total del tratamiento hasta 150.000, refleja la implementación de protocolos altamente personalizados. Esta personalización terapéutica, aunque clínicamente justificada, genera un impacto económico considerable con una media de costes promedio de 42.267 euros por paciente, donde la quimioterapia representa el principal factor de variabilidad financiera.

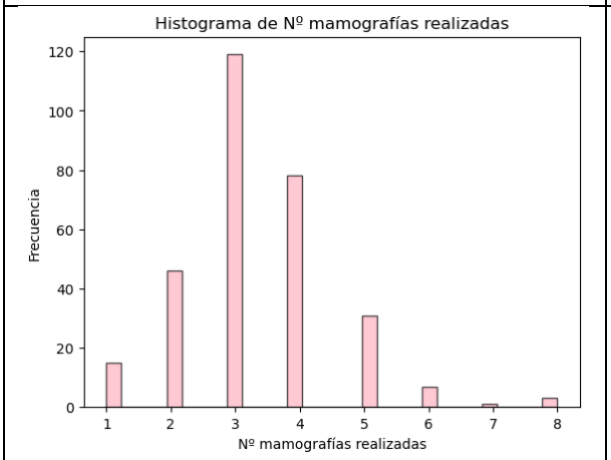


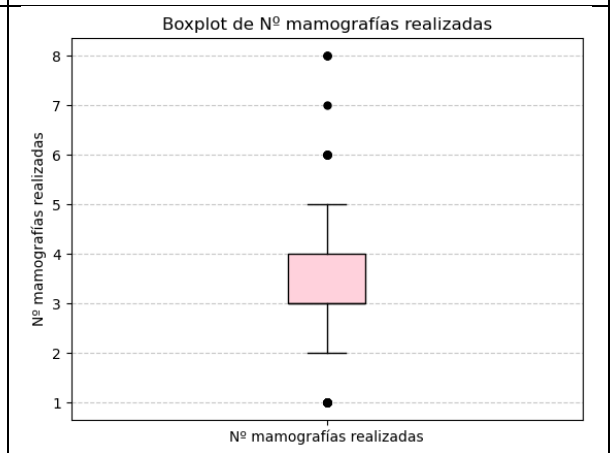

Histograma de Estadio
Frecuencia
Estadio

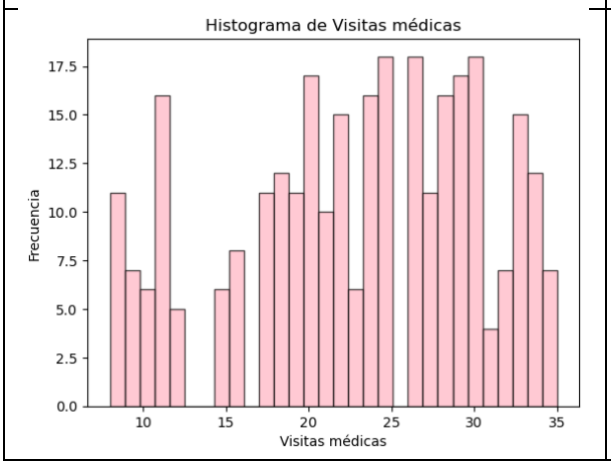

Boxplot de Estadio
Estadio
Estadio

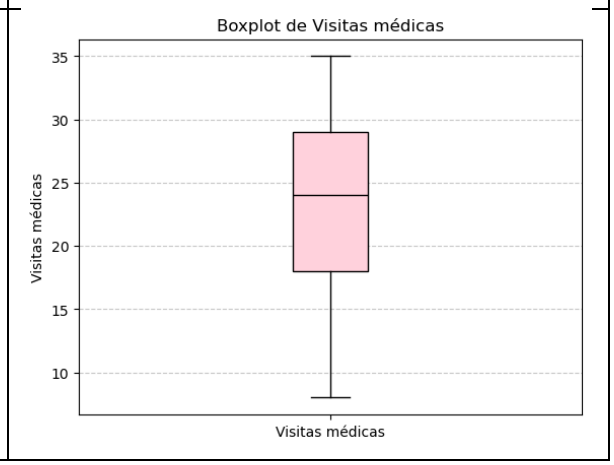

Histograma de Ganglios axilares
Frecuencia
Ganglios axilares

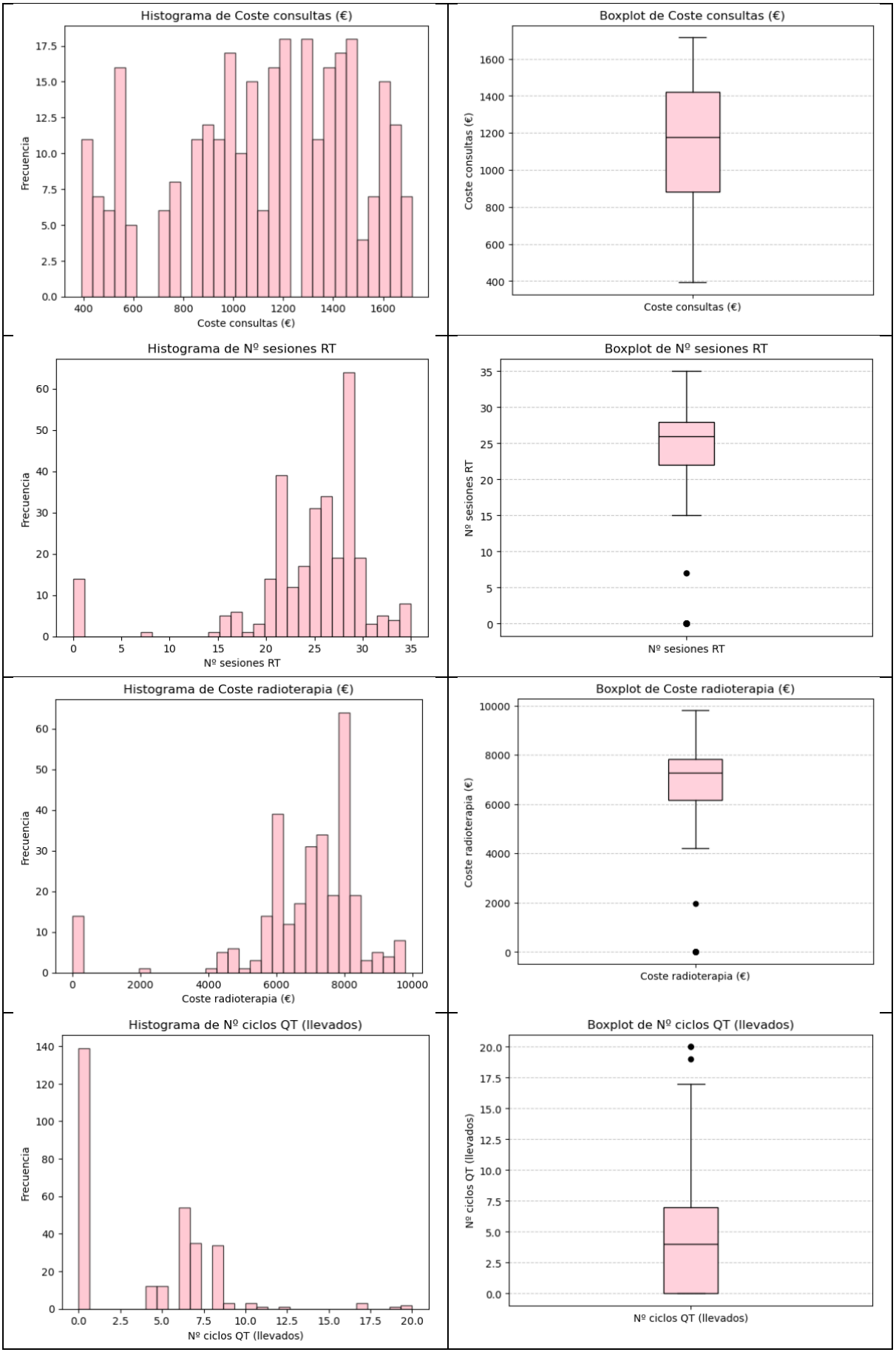

Boxplot de Ganglios axilares
Ganglios axilares
Ganglios axilares


Histograma de Nº mamografías realizadas
Frecuencia
Nº mamografías realizadas


Boxplot de Nº mamografías realizadas
Nº mamografías realizadas
Nº mamografías realizadas


Histograma de Visitas médicas
Frecuencia
Visitas médicas


Boxplot de Visitas médicas
Visitas médicas
Visitas médicas


Histograma de Coste consultas (€)
Frecuencia
Coste consultas (€)
Boxplot de Coste consultas (€)
Coste consultas (€)
Histograma de Nº sesiones RT
Frecuencia
Nº sesiones RT
Boxplot de Nº sesiones RT
Nº sesiones RT
Histograma de Coste radioterapia (€)
Frecuencia
Coste radioterapia (€)
Boxplot de Coste radioterapia (€)
Coste radioterapia (€)
Histograma de Nº ciclos QT (llevados)
Frecuencia
Nº ciclos QT (llevados)
Boxplot de Nº ciclos QT (llevados)
Nº ciclos QT (llevados)

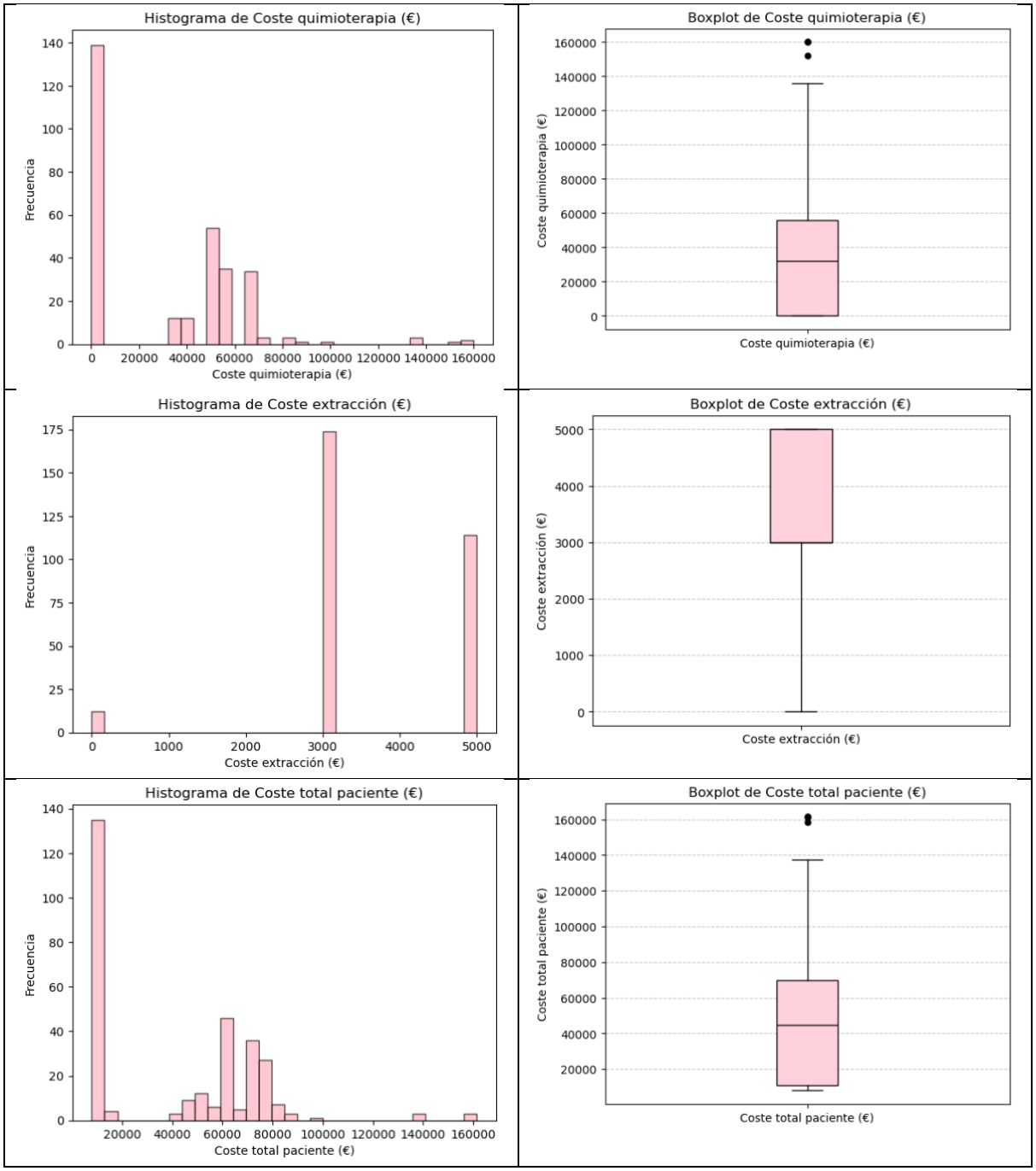


Tabla 29 Gráficos exploratorios del modelo de Regresión

Fuente: Elaboración propia con Python

De los siguientes gráficos se puede extraer que la variable edad dibuja una distribución prácticamente uniforme entre los veinte y los noventa años, lo que indica que el estudio no está sesgado hacia pacientes especialmente jóvenes ni mayores. El estadio tumoral, en cambio, se concentra de forma muy visible en los niveles I y II, con una proporción mucho menor de casos avanzados. El tamaño tumoral, aunque abarca desde un centímetro hasta algo por encima de veinticinco, revela la mayor densidad entre los dos y cinco centímetros y se inclina ligeramente hacia tamaños superiores, manteniendo de nuevo esa asimetría positiva.

En las variables clínicas aparece un patrón similar. Las visitas médicas oscilan entre ocho y treinta y cinco, con un pico consistente alrededor de las treinta, mientras que las mamografías se acumulan en torno a tres o cuatro por paciente. El número de sesiones de radioterapia se sitúa mayoritariamente entre veintidós y treinta, y la quimioterapia, por su lado, forma dos grupos diferenciados: un grupo numeroso de pacientes no recibe ningún ciclo y otro completa entre cinco y ocho. Los costes unitarios permanecen constantes; no añaden variabilidad por sí mismos, pero al multiplicarse por la utilización amplifican las diferencias de gasto total. Este último muestra dos clústeres como el caso de la quimioterapia muy definidos: uno "bajo", cercano a los diez-quince mil euros, y otro "alto", en torno a los sesenta-ochenta mil, con unos pocos valores extremos.

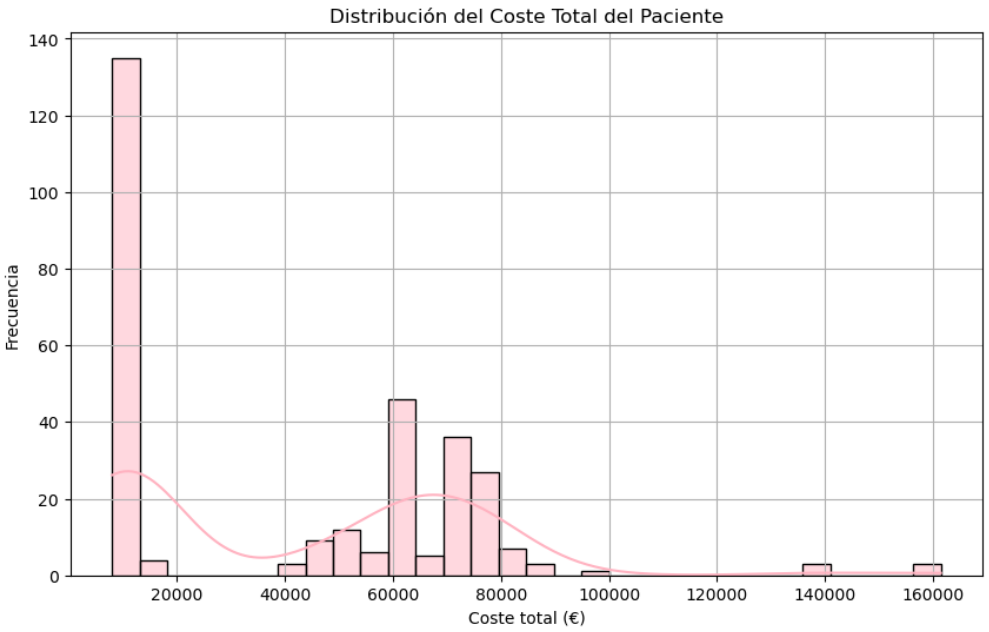


Tabla 30 Distribución de los costes

Fuente: Elaboración propia con Python

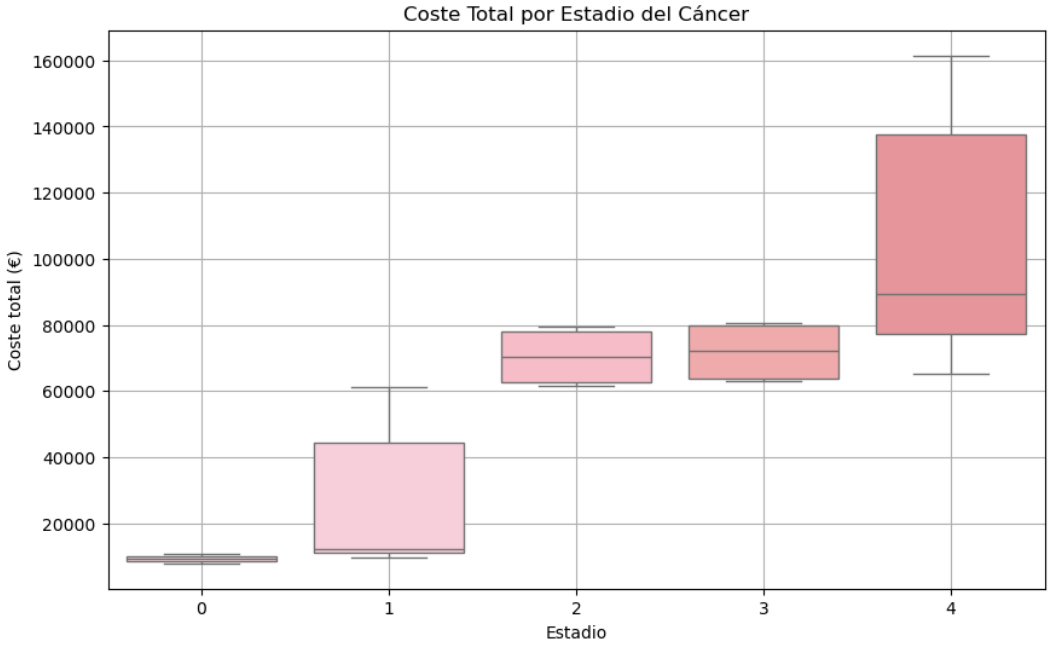


Tabla 31 BoxPlot de la distribución del coste por estadio

Fuente: Elaboración propia con Python

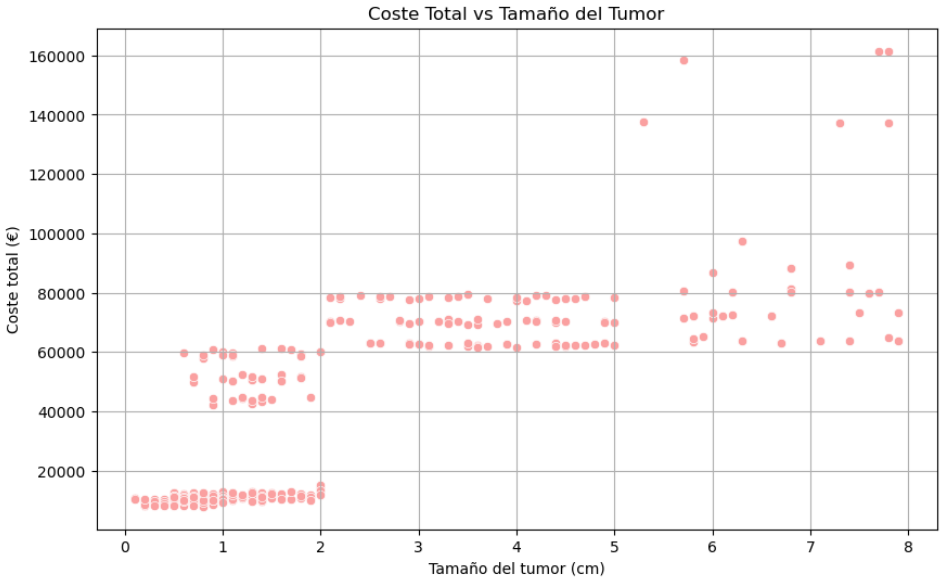


Tabla 32 Coste frente al tamaño del tumor en cm

Fuente: Elaboración propia con Python

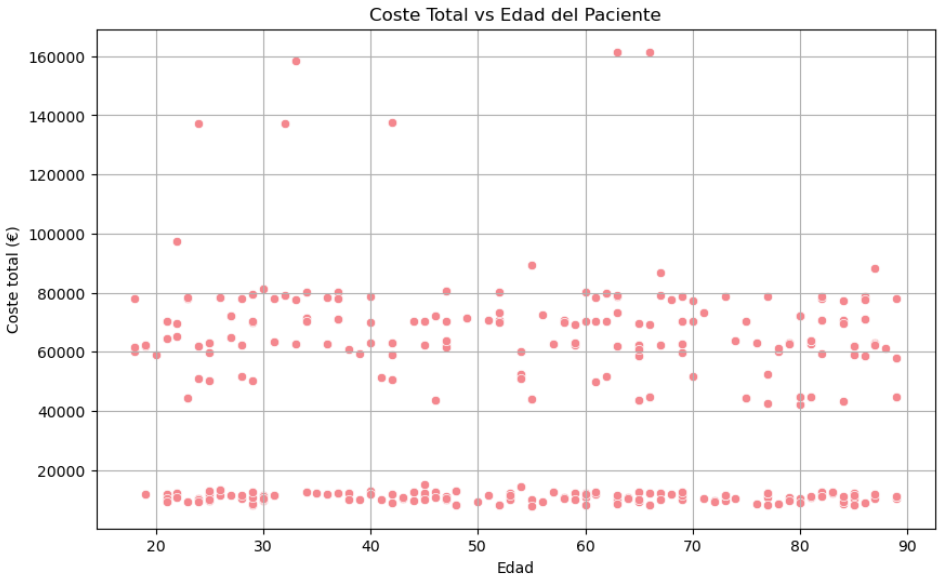


Tabla 33 Coste frente a la edad del paciente

Fuente: Elaboración propia con Python

Pasando al análisis bivariante, los boxplot evidencian que el estadio del tumor es el principal motor económico: tanto la mediana como la dispersión del coste total aumentan de forma casi monótona al pasar de estadio 0 a estadio IV, y el rango intercuartílico se ensancha visiblemente en los casos más avanzados. Por el contrario, la nube de puntos que relaciona coste y tamaño tumoral apenas muestra pendiente; los gastos elevados y moderados conviven a lo largo de todo el eje horizontal, señal de que el tamaño, por sí solo, explica muy poco. Algo parecido sucede con la edad: los pacientes caros y los baratos se reparten de manera casi uniforme entre los veinte y los noventa años, lo que sugiere que la variable demográfica carece de poder explicativo directo.

En conclusión, el nivel de costes está impulsado sobre todo por la intensidad terapéutica y por la fase de la enfermedad, mientras que factores como la edad o incluso el propio tamaño del tumor resultan secundarios. Un reducido subgrupo de pacientes en estadios avanzados concentra la mayor parte del presupuesto. Desde el punto de vista la gestión hospitalaria, las estrategias de contención deberían dirigirse a optimizar la indicación y duración de los tratamientos costosos, sobre todo quimioterapia y radioterapia, en esos estadios III y IV.

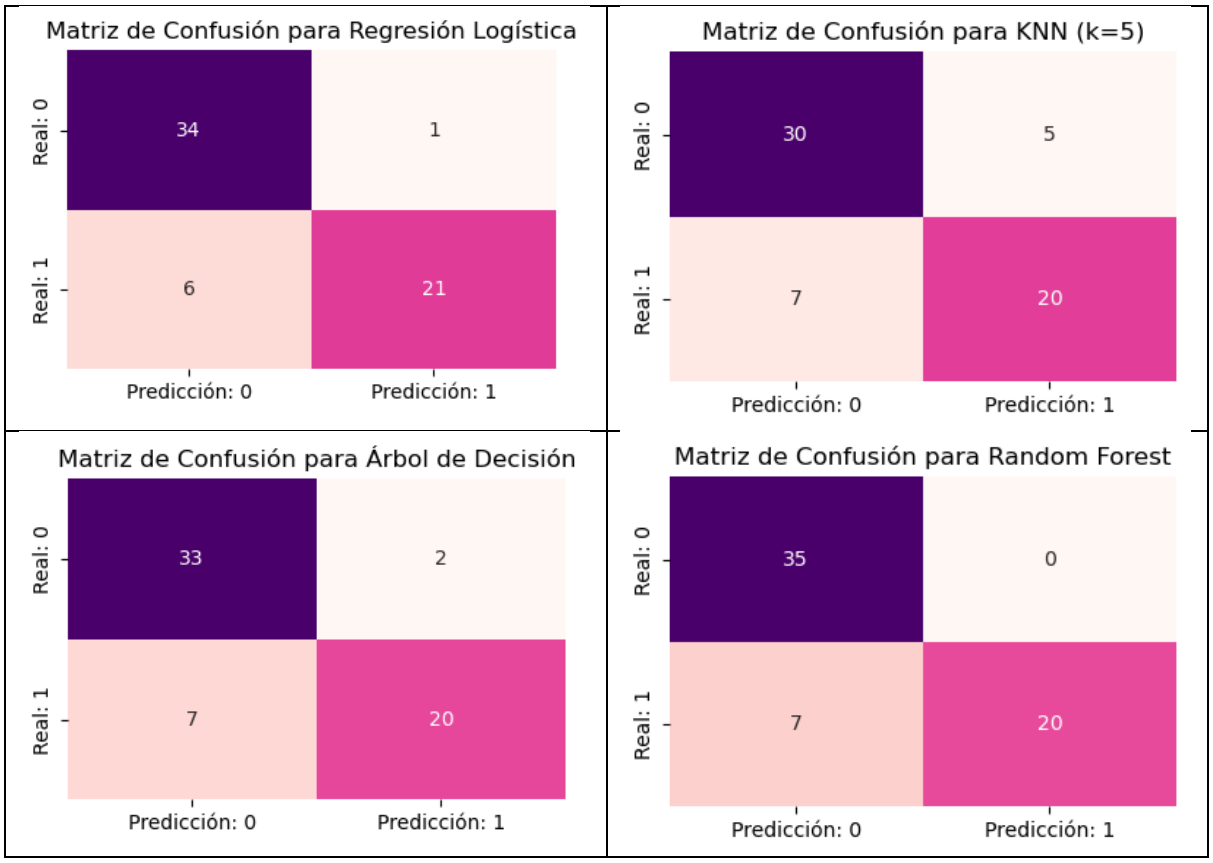
5.2 Métodos de clasificación

En el marco de la detección temprana del cáncer de mama, se ha llevado a cabo un análisis exhaustivo cuyo objetivo principal es identificar el algoritmo de *machine learning* capaz de ofrecer el equilibrio óptimo entre sensibilidad y precisión, es decir, que el algoritmo detecte correctamente los tumores malignos y a simultáneamente reduzca la probabilidad de que un resultado positivo sea realmente maligno. Para llevar a cabo el objetivo, ello se entrenaron seis modelos clásicos sobre la base de datos explicada en el punto 4.1.1 y tras haber sido tratada en el analisis preliminar descrito en el punto 5.1 de este documento. Una vez depurada, la variable objetivo a predecir será Diagnosis *Result* la cual toma el valor de 0 si el tumor es benigno y en el caso contrario adoptará el valor de 1.

Con el fin de asegurar estimaciones imparciales del rendimiento, el conjunto de datos se dividió mediante una partición estratificada: el 70 % de las observaciones se destinó al entrenamiento y el 30 % restante a la fase de prueba, manteniéndose en ambos subconjuntos la proporción original de casos benignos y malignos. Todas las variables numéricas se estandarizaron (media cero y desviación típica uno) dentro de una *pipeline*, de modo que la transformación no «viera» nunca los datos reservados para la evaluación.

Se evaluaron seis algoritmos clásicos: Regresión Logística, Random Forest, Gradient Boosting, Árbol de Decisión, *k*-Nearest Neighbors (con $k = 5$) y una Máquina de Vectores de Soporte (SVM) con núcleo RBF. El desempeño de cada modelo se midió mediante cinco indicadores: *accuracy*, precisión (valor predictivo positivo), sensibilidad (*recall*), F1-score y área bajo la curva ROC (AUC). En un programa de cribado poblacional la sensibilidad reviste especial importancia, dado que un falso negativo implica la posibilidad de omitir un cáncer en estadio tratable.

A continuación, se muestran los gráficos que muestran la matriz de confusión de cada uno de los modelos a analizar.



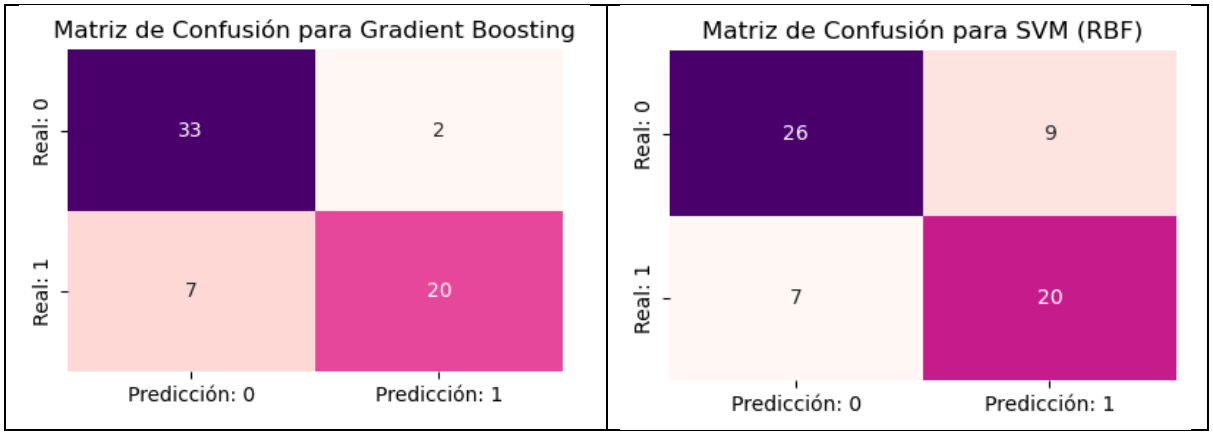


Tabla 34 Matrices de confusión del Modelo de clasificación

Fuente: Elaboración propia con Python

Los resultados muestran matices relevantes. La Regresión Logística alcanzó la sensibilidad más alta (0,778) con un F1-score de 0,857 y un *accuracy* de 0,887, lo que la convierte en la alternativa más segura cuando la prioridad es detectar el mayor número posible de tumores. El modelo Random Forest, por su parte, obtuvo la máxima precisión (1,000) y la AUC más amplia (0,941), aunque su sensibilidad (0,741) quedó ligeramente por debajo de la de la Logística. Gradient Boosting y el Árbol de Decisión presentaron un equilibrio intermedio (precisión y sensibilidad de 0,909 y 0,741, respectivamente), si bien el árbol individual exhibió indicios de sobreajuste, reflejados en una AUC ligeramente inferior (0,842). Finalmente, *k*-Nearest Neighbors y la SVM mostraron la actuación más discreta, lastrados, respectivamente, por la alta dimensionalidad y por la falta de optimización de hiperparámetros.

Modelo	Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest	0,8871	1,0000	0,7407	0,8511	0,9413
Gradient Boosting	0,8548	0,9091	0,7407	0,8163	0,9386
KNN (k=5)	0,8065	0,8000	0,7407	0,7692	0,9063
Regresión Logí-stica	0,8871	0,9545	0,7778	0,8571	0,9058
SVM (RBF)	0,7419	0,6897	0,7407	0,7143	0,8847
Árbol de Decisión	0,8548	0,9091	0,7407	0,8163	0,8418

Tabla 35 Resultados del modelo de clasificación

Fuente: Elaboración propia con Python

De igual modo, la interpretabilidad es un requisito para la adopción clínica. La Regresión Logística permite vincular cada coeficiente con un factor de riesgo tangible, mientras que existen técnicas que pueden revelar qué variables impulsan la decisión en un Random Forest individualizado para cada paciente. Por tanto, ambos modelos son los algoritmos con mejores métricas; cualquiera de ellos podría resolver de forma óptima el problema inicial.

5.3 Métodos de predicción

En el contexto actual de la gestión sanitaria, la predicción precisa de costes de tratamiento representa un desafío fundamental que impacta directamente en múltiples dimensiones

operacionales y estratégicas. La sostenibilidad financiera de las instituciones sanitarias depende en gran medida de la capacidad de anticipar y planificar los recursos necesarios para cada paciente. Simultáneamente, la calidad de atención al paciente se ve influenciada por la disponibilidad adecuada de recursos, mientras que la eficiencia operacional de los servicios médicos requiere una coordinación precisa entre demanda y capacidad instalada.

Es por eso por lo que, con el fin de predecir el coste total asociado al tratamiento del cáncer de mama, se han desarrollado varios algoritmos de regresión y así encontrar el más adecuado a dicho objetivo. Para ello, se ha utilizado la base de datos explicada en el punto 4.1.2.1. la cual comprende la información completa de 300 pacientes con cáncer de mama.

Antes de entrenar los algoritmos, Los datos se dividieron estratégicamente en un conjunto de entrenamiento que representa el 80% de las observaciones, equivalente a 240 casos, y un conjunto de prueba que constituye el 20% restante, correspondiente a 60 observaciones. Esta división se realizó mediante un método de selección aleatoria estratificada que garantiza la representatividad de ambos conjuntos y minimiza el sesgo de selección.

Tras esta división, se desestimó la opción de entrenar la regresión lineal simple debido a que el coste total del tratamiento de un paciente con cáncer no se relaciona linealmente con una sola variable, depende de diversos factores.

Respecto a los modelos restantes vistos en el punto 3.2.2, se entrenó un modelo de regresión lineal múltiple teniendo en cuenta todas las variables del dataset. Los resultados de las métricas de evaluación sobre la regresión lineal múltiple muestran un RMSE del 2,60E-05 y un R2 de 1. Esto sugiere un sobreajuste en los datos debido a que el dataset contiene todo el desglose de los costes unitarios y totales por tratamiento realizado y por tanto el modelo aprende fácilmente a predecir el coste total. Por este motivo, se decide volver a entrenar la regresión lineal múltiple utilizando las variables clínicas como *Edad, Años con la enfermedad, Estadio, Tumor (cm), Ganglios axilares, Nº mamografías realizadas* y *Tipo de cáncer*, además de la edad del paciente y la ciudad de residencia.

Una vez eliminadas las variables de coste, el valor del R2 se reduce a un 0,5787 y el RMSE aumenta a 22.850,11, no obstante, ahora el R2 obtenido no resulta óptimo para predecir el coste total que implica la enfermedad para el paciente. A continuación, dado que la regresión de Ridge se utiliza específicamente para evaluar el impacto de la regularización L2 en el rendimiento predictivo, incorporando técnicas para controlar el sobreajuste potencial, se prueba obtener de nuevo los resultados de la regresión lineal múltiple con el objetivo de ver cuanto mejora el R2.

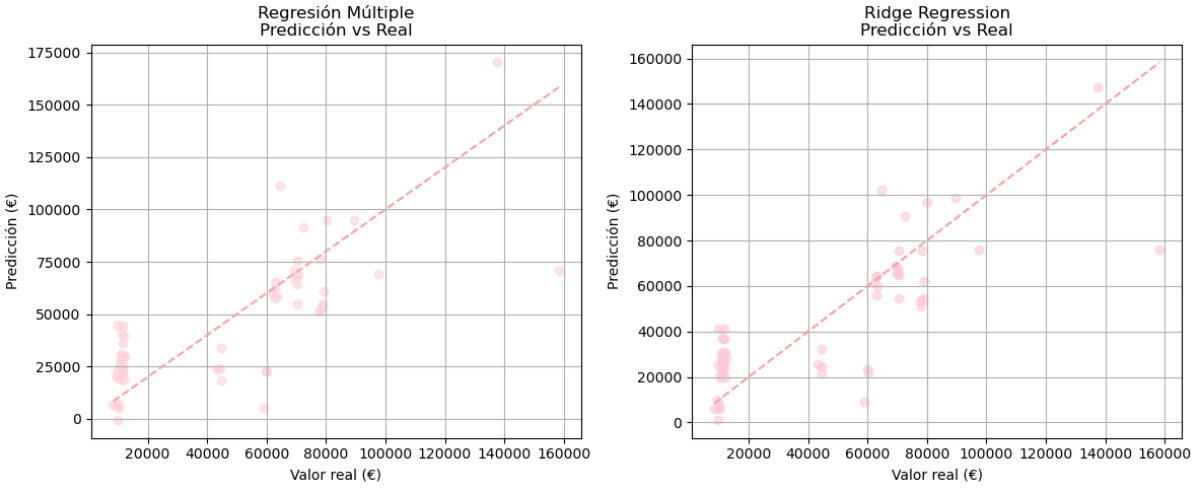


Tabla 36 Regresión múltiple con o sin ajuste

Fuente: Elaboración propia con Python

A pesar de que se consigue una ligera mejoría obteniendo un RMSE de 21.189,15 y un R^2 de 0,6377, indicando que la regularización L2 proporciona beneficios apreciables en la precisión predictiva, debido a que el problema a tratar va ligado a un tema tan delicado como el coste sanitario, se plantea la posibilidad de probar un modelo diseñado específicamente para capturar interacciones no lineales complejas entre las variables predictoras como es el Random Forest Regressor.

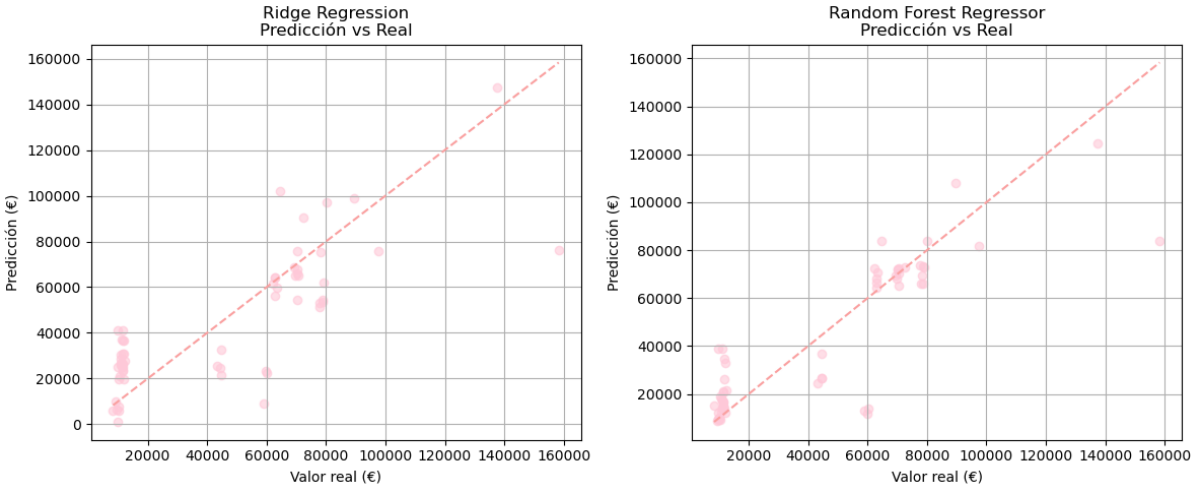


Tabla 37 Rigde y Random forest con ajustes

Fuente: Elaboración propia con Python

El modelo Random Forest superó notablemente a ambos modelos lineales en todas las métricas evaluadas. Su capacidad para reducir el RMSE hasta 17,491.42 € representa una disminución de 5,358.69 € en comparación con la Regresión Lineal y de 3,697.73 € respecto a la Regresión Ridge. De la misma manera, el aumento del R^2 hasta 0,75 demuestra que este modelo puede explicar el 75% de la variabilidad en los costes totales, superando en 17 puntos porcentuales a la Regresión Lineal y en 11 puntos porcentuales a la Regresión Ridge.

Desde una perspectiva clínica y económica, las diferencias observadas tienen implicaciones prácticas significativas. La reducción de más de 5,000 € en el error de predicción promedio entre Random Forest y Regresión Lineal representa una mejora en la capacidad de planificación financiera. Esta precisión adicional permite a los gestores sanitarios realizar estimaciones presupuestarias más confiables y optimizar la asignación de recursos disponibles.

La capacidad del Random Forest para explicar un 17% adicional de la variabilidad en los costes, comparado con la Regresión Lineal, indica que factores previamente considerados como aleatorios o impredecibles pueden ahora ser incorporados en las predicciones lo cual proporciona una comprensión más completa y precisa de los determinantes económicos del tratamiento oncológico.

5.4 Limitaciones y líneas de mejora para los modelos utilizados.

El cáncer de mama representa uno de los tipos de cáncer más prevalentes a nivel mundial, siendo crucial el desarrollo de herramientas de diagnóstico asistido que permitan una clasificación precisa y temprana. La aplicación de técnicas de machine learning en el análisis de datos clínicos ofrece oportunidades significativas para mejorar la precisión diagnóstica y optimizar los protocolos de tratamiento. No obstante, el uso de estos algoritmos tiene sus limitaciones.

Una de las principales limitaciones de este trabajo en concreto es el dataset utilizado ya que el tamaño muestral de 213 observaciones, aunque adecuado para el análisis exploratorio, puede presentar limitaciones para la generalización a poblaciones más amplias. También cabe mencionar que la posible existencia de variables predictoras adicionales no incluidas en el dataset actual podría limitar el poder predictivo máximo alcanzable.

Teniendo en cuenta la confidencialidad de los datos clínicos de los pacientes, el objetivo del trabajo es presentar estrategias útiles para la detección temprana de dicha enfermedad explicando su interpretabilidad en un posterior uso con datos reales.

En cuanto a aspectos de mejora en el caso de clasificación, para que el modelo sea útil en la práctica médica, los doctores deben poder entender cómo tomar sus decisiones y confiar en sus resultados. También es importante probar el modelo con diferentes grupos de pacientes para asegurar que funciona igual de bien independientemente de la edad, origen u otras características personales. Siempre que sea posible, el modelo debe probarse con datos completamente nuevos de otros hospitales o estudios para confirmar que funciona correctamente fuera del grupo de pacientes original con el que fue entrenado.

En el caso de los modelos de regresión, la investigación surge de la necesidad crítica de optimizar la planificación de recursos sanitarios y mejorar la gestión presupuestaria en el sector salud, especialmente en un contexto donde la precisión en la estimación de costes determina la sostenibilidad financiera de las instituciones sanitarias.

Siguiendo con las casuísticas respecto el dataset, en este caso también se encuentran limitaciones con los datos. Se trata de un conjunto de datos simulados, aunque estén basados en datos robustos de los informes, datos clínicos, y la estadificación TNM. Del mismo modo que en clasificación, el número de

pacientes y la no disponibilidad consistente de las variables predictivas utilizadas en el entorno clínico real afecta al rendimiento y la interpretación precisa del modelo de regresión.

Con la intención de poder mejorar la optimización de los costes, se propone el desarrollo de un sistema de actualización automática que permita al modelo adaptarse cuando cambien los protocolos de tratamiento o se introduzcan nuevas tecnologías médicas. Esto garantizará que las predicciones sigan siendo precisas incluso cuando evolucionen las prácticas clínicas.

6. Dashboard

6.1 Explicación y desarrollo del dashboard

La digitalización y el uso de herramientas de Business Intelligence han comenzado a consolidarse como parte de los procesos estratégicos de muchas organizaciones. No obstante, no todas las empresas o sectores pueden avanzar al mismo ritmo, y particularmente en el ámbito sanitario, se encuentran con limitaciones presupuestarias y estructurales que dificultan en ocasiones la inversión en investigación, desarrollo e innovación (I+D+i).

Aun así, la implementación de soluciones como el uso de los dashboards interactivos es una solución accesible y escalable para optimizar la toma de las decisiones tanto a nivel económico como a nivel clínico. En este caso concreto se ha desarrollado un dashboard interactivo con la finalidad de demostrar cómo la visualización de los datos puede apoyar y desarrollar la eficiencia de los programas de detección precoz además no solo permite poder llevar un control de los pacientes, sino que además permite entender el coste económico promoviendo que se cree una cultura organizacional orientada a la gestión basada en datos y más en entornos como la salud que presenta recursos limitados.

La herramienta con la que se ha decidido elaborar este dashboard ha sido PowerBI, en este caso se ha decidido plantear un dashboard tipo que sintetiza de una forma cómoda e interactiva los principales indicadores extraídos de los conjuntos de datos simulados. La idea principal es facilitar la exploración visual de tendencias clínicas y económicas asociadas al cáncer de mama.

La estructura del panel está diseñada con la finalidad de ofrecer información útil de forma clara, precisa, rápida e intuitiva a distintos perfiles profesionales: gestores, médicos, responsables públicos o investigadores. La propuesta se basa a partir de los datos simulados crear una representación de cómo sería el panel final, en este caso orientado a un profesional sanitario, lo ideal sería vincularlo con las bases de datos del hospital y elaborar dashboards personalizados para cada perfil profesional.

La distribución visual ofrece una lectura rápida y un análisis estratégico. Para ello se han empleado tarjetas de resumen (KPI) para métricas globales como el coste total, la edad media, los años promedio con la enfermedad y el número de pacientes, estas métricas permiten un diagnóstico inmediato del alcance del problema.

Tras ello se han representado diagramas de sectores para variables clínicas (metástasis, tipo de cirugía, estadio), histogramas de frecuencia, y una visualización cartográfica que permite identificar patrones geográficos y desigualdades regionales para poder comprender mejor el gasto que destina cada comunidad autónoma.

Los filtros interactivos ubicados en la parte izquierda del panel permiten segmentar la información por ciudad, paciente y fecha, facilitando un análisis a medida. Esta funcionalidad aporta valor a la toma de decisiones porque permite observar diferencias entre poblaciones, detectar ineficiencias en ciertos territorios y aplicar intervenciones dirigidas. Además, la selección de indicadores fue realizada considerando criterios de relevancia clínica, impacto económico y utilidad política, lo que confiere al dashboard un enfoque práctico para contextos de planificación sanitaria. Además, se encuentra

personalizado ofreciendo de forma individual un: Bienvenida doctora, junto con una imagen de archivo.

El dashboard también incorpora secciones específicas sobre el total de procedimientos realizados (número de biopsias, mamografías, visitas médicas, ciclos de quimioterapia y sesiones de radioterapia), así como el tamaño promedio tumoral, lo que permite una valoración agregada de la carga asistencial. Estas visualizaciones facilitan el análisis de los costes asociados, aportando información clave para estudios de coste-efectividad y optimización de recursos.

Desde el punto de vista técnico y estratégico, la estructura visual se ha sido diseñada para maximizar la legibilidad y favorecer una navegación fluida. Los colores suaves con tonos rosados que siguen la línea de colores que se asocian al cáncer de mama, la jerarquía tipográfica clara y la disposición modular de los elementos que permiten en todo momento al usuario identificar rápidamente las áreas más importantes.

En los anexos se encuentra el enlace de acceso al dashboard interactivo.

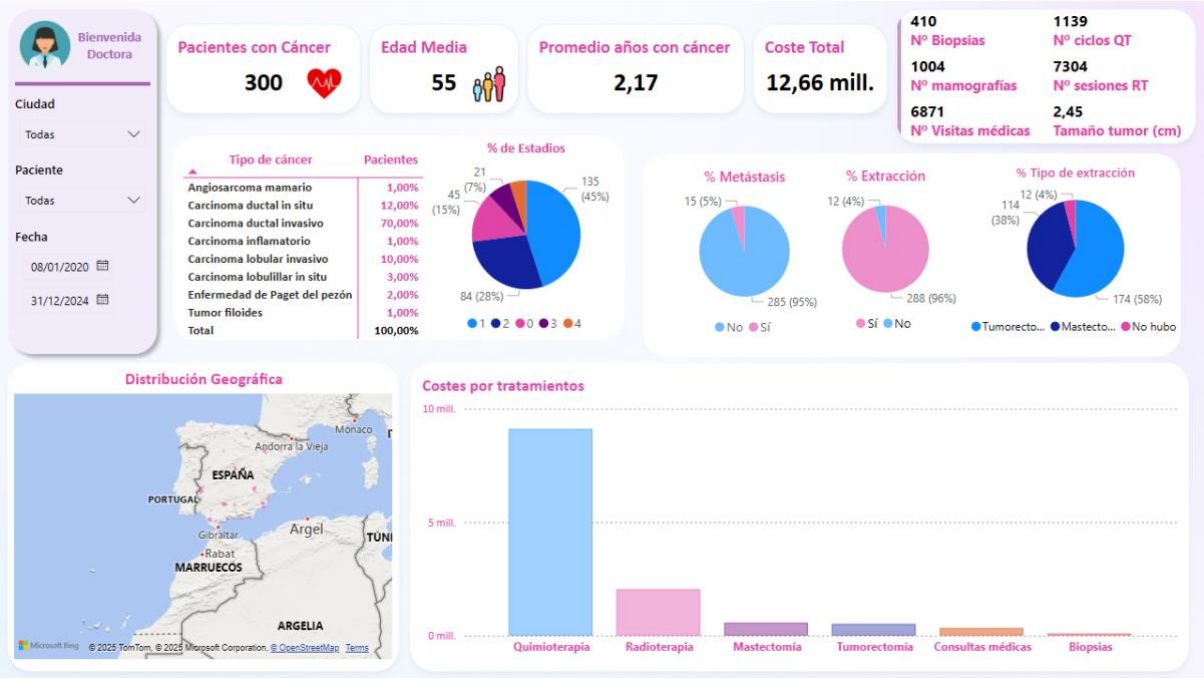


Figura 1.42. Dashboard clínico
Fuente: Elaboración propia con PowerBI

6.2 Limitaciones y propuestas de mejoras

Como ya se ha mencionado implementar un dashboard ofrece multiples ventajas, además de ser una herramienta de apoyo a la hora de tomar decisiones en el ámbito sanitario, es importante reconocer ciertas limitaciones que afectan tanto a su aplicabilidad como a la interpretación de los resultados obtenidos. En este punto se va a reflexionar sobre los aspectos que pueden mejorarse cara a los futuros desarrollos, así como proponer líneas de acción para su evolución.

Una de las principales limitaciones y de la más evidente es en el origen de los datos utilizados, en este caso se trata de un conjunto de datos simulados que no se cuenta con la robustez ni representatividad de un dataset real procedente de instituciones sanitarias, aunque se ha buscado y se ha asegurado de reflejar las distribuciones realistas basadas en los informes, datos clínicos, y la estadificación TNM. En futuras investigaciones, sería deseable acceder a bases de datos hospitalarias reales, con los correspondientes permisos éticos y de privacidad, para validar y refinar los modelos propuestos.

Otro aspecto a considerar es que la herramienta ha sido diseñada desde una perspectiva generalista orientada al personal sanitario, sin haber desarrollado versiones específicas por perfil profesional como pueden ser la administración pública tanto a nivel económico como a nivel clínico, los oncólogos o los administradores de los hospitales (el equipo de contabilidad).La idea es personalizar según las necesidades específicas de cada tipo de usuario, ofreciendo indicadores e interfaces adaptadas a sus competencias y objetivos.

También se observa una limitación en la temporalidad de los datos, puesto que actualmente no ofrece más que ciertas métricas temporales, la idea sería incorporar más gráficos que puedan mostrar la evolución de la enfermedad en términos de tiempo, como pueden ser las series temporales o la evolución longitudinal de pacientes. Todo esto permite analizar la progresión de la enfermedad, la respuesta a los tratamientos y la evolución de los costes a lo largo del tiempo. Integrar esta lógica temporal supondría una mejora sustancial para la planificación sanitaria y el seguimiento personalizado.

Desde el punto de vista técnico, el panel actual de PowerBI presenta una visualización estática de datos que limita la automatización del sistema. Una mejora estratégica sería su integración directa con sistemas de información hospitalaria, permitiendo actualizaciones en tiempo real y la obtención de alertas automáticas ante umbrales clínicos o financieros.

Finalmente, y tras la elaboración de los modelos de clasificación y regresión se plantea la incorporación de algoritmos de machine learning que permiten generar predicciones o clasificaciones automáticas sobre los datos que haya en la base a los datos. La futura integración de modelos predictivos podría enriquecer la funcionalidad del dashboard, facilitando la detección de pacientes de alto riesgo, el número de sesiones o la estimación de los costes esperados según variables individuales.

Aunque el dashboard actual constituye una herramienta valiosa desde el punto de vista exploratorio y comunicativo, su perfeccionamiento implica la incorporación de datos reales, personalización por perfiles, capacidad de análisis temporal, conexión con sistemas clínicos y aplicación de técnicas predictivas. Estos avances permitirán transformar el dashboard en una herramienta integral, dinámica y orientada a la mejora continua de la gestión del cáncer de mama en el sistema sanitario.

7. Conclusiones

Durante este Trabajo Fin de Máster se ha abordado de forma integral la problemática del cáncer de mama desde una perspectiva multidisciplinar, combinando los datos clínicos, las herramientas de Business Intelligence (BI) y los criterios económicos para demostrar la importancia que tiene la detección precoz en la sostenibilidad del sistema sanitario, ya que detectarlo y tratar un cáncer en una fase inicial se traduce en un reducir el coste hasta cinco veces menos que en fases avanzadas.

El objetivo principal del se ha basado en analizar cómo el uso de la inteligencia de negocios puede contribuir a la mejora de los programas de cribado mamográfico y optimizar el gasto público asociado. Para alcanzar este fin, se plantearon varios objetivos específicos como fueron analizar la relación entre el estadio de diagnóstico y los costes sanitarios, construir modelos predictivos mediante clasificación y regresión con base en datos simulados representativos del contexto español y desarrollar un dashboard interactivo orientado a la visualización de indicadores clave para distintos perfiles profesionales del sistema de salud.

Estos objetivos han sido abordados y resueltos a lo largo de los apartados 4, 5 y 6 de este trabajo. En primer lugar, se ha demostrado que los costes sanitarios derivados del tratamiento del cáncer de mama aumentan de forma sustancial a medida que el diagnóstico se produce en estadios más avanzados. Esta afirmación se sustenta en la recopilación y análisis de literatura nacional e internacional, así como en los datos simulados construidos en base a guías clínicas como el sistema TNM y fuentes como el Ministerio de Sanidad (2023).

En segundo lugar, se han desarrollado modelos de clasificación utilizando algoritmos como Random Forest, KNN, SVM, árbol de decisión y Regresión Logística, y modelos de regresión mediante una regresión múltiple, regresión múltiple con Ridge y random forest regressor , cuyos resultados han sido documentados en el apartado 5.2 y 5.3.

En particular, los modelos de clasificación basados en Random forest y Gradient Boosting ofrecieron un rendimiento superior, con buena capacidad explicativa, permitiendo discriminar entre tumores benignos y malignos y estimar con precisión los costes económicos asociados a cada caso clínico, véase en la tabla 35. Por tanto, en conclusión, si la prioridad es no pasar por alto un solo tumor maligno, la Regresión Logística se perfila como la alternativa preferente gracias a su sensibilidad y fácil interpretación. Ahora bien, cuando el entorno penaliza severamente los falsos positivos, por su coste económico o psicológico, la precisión perfecta y la mayor capacidad discriminatoria del Random Forest justifican su elección.

En el modelo de regresión finalmente se opta por el modelo de Radom Forest Regressor que consigue añadir un 17% de capacidad explicativa de los costes sanitarios, siendo esto una mejora en la planificación financiera de los gestores sanitarios tal y como se explica en el apartado 5.3. Además, gracias a estos modelos podemos anticipar los costes de los tratamientos, reduciendo miles de euros. De esta manera se podría integrar los resultados con los sistemas de compras de insumos para ajustar presupuestos.

El último objetivo se ha cumplido mediante la elaboración de un dashboard interactivo en Power BI, en el que se integran de forma visual los indicadores clínicos, geográficos y económicos. Esta

herramienta permite una segmentación dinámica por ciudad, estadio clínico, tipo de intervención y perfil de paciente. Entre las funcionalidades destacan la estimación de coste total acumulado (12,68 millones de euros en el dataset simulado), el desglose de procedimientos como mamografías, biopsias y tratamientos aplicados, un mapa de calor interactivo que permite observar desigualdades territoriales en la carga asistencial y un diagrama de coste por tipología.

Los resultados obtenidos, tanto en el análisis teórico como en las simulaciones aplicadas, muestran que el coste del tratamiento varía considerablemente según el estadio en el que se detecta la enfermedad, lo cual confirma la hipótesis principal de que es necesario que se actúe a tiempo ya que ayuda a salvar vidas y además también optimiza los recursos públicos.

No obstante, el trabajo presenta limitaciones, entre las cuales destaca el uso de bases de datos simuladas, que, si bien han sido diseñadas siguiendo criterios clínicos realistas, requieren validación empírica con datos reales para su aplicación directa. Asimismo, el modelo se presenta como una propuesta tipo, sin conexión directa con sistemas de información hospitalarios. Esta circunstancia abre oportunidades futuras: vincular los dashboards con fuentes reales, añadir un mayor volumen de datos médicos de cada paciente en el modelo de clasificación, incorporar modelos predictivos más avanzados, y personalizar paneles por tipo de usuario (clínico, gestor, planificador).

Desde una perspectiva organizativa, se destaca que una de las oportunidades más relevantes es anticipar las necesidades en recursos humanos, en especial en perfiles clave como la enfermería, altamente demandados en la atención oncológica. La capacidad de prever la carga asistencial según el estadio clínico permitiría distribuir mejor el personal, reducir horas extra innecesarias y mejorar la calidad asistencial. Esta optimización de la carga de trabajo no solo eleva la calidad de vida del paciente, sino también la del profesional sanitario.

Desde una perspectiva de gobernanza a nivel sanitaria, se plantea incentivar la transparencia y la mejora continua mediante la publicación periódica de indicadores comparativos entre hospitales o comunidades autónomas. Publicar trimestralmente rankings de costes medios por pacientes, tiempos de diagnóstico y eficacia de tratamiento podría fomentar una competencia sana y estímulos para la mejora continua. Es primordial seguir invirtiendo en desarrollo e innovación ya que como se puede comprobar es fundamental para mejorar y salvar las vidas de los pacientes y a su vez una inversión actual permitirá que a largo plazo pueda haber menos costes por tratamientos.

Además, se ha conseguido vincular de manera explícita este trabajo con los Objetivos de Desarrollo Sostenible (ODS), especialmente en lo relativo al ODS 3 (Salud y Bienestar), el ODS 5 (Igualdad de Género), el ODS 9 (Industria, Innovación e Infraestructura) y el ODS 17 (Alianzas para lograr los objetivos). El enfoque adoptado demuestra que el uso de BI en salud no solo es técnicamente viable, sino que representa una herramienta de alto valor estratégico en políticas públicas modernas orientadas a la equidad, la eficiencia y la innovación, como se argumenta en el apartado de los ODS del documento.

En resumen y para finalizar, este TFM remarca la importancia de aplicar la inteligencia de negocios al ámbito sanitario ya que puede transformar la forma en que se planifican, evalúan y ejecutan estrategias de detección precoz. Constituye un ejemplo práctico y replicable de cómo el análisis de datos puede ser una herramienta de alto impacto en la toma de decisiones clínicas y económicas, en línea con los desafíos actuales de la salud pública global. El conocimiento generado ofrece una base

sólida sobre la que seguir construyendo soluciones tecnológicas aplicadas a problemas sociales complejos, como lo es el cáncer de mama.

8. Bibliografía

Agencia Española de Medicamentos y Productos Sanitarios. (2021). *Informe de posicionamiento terapéutico de talazoparib (Talzenna®) en pacientes con cáncer de mama HER-2 negativo con mutaciones BRCA 1/2 en progresión a tratamientos previos* [Informe público]. https://www.aemps.gob.es/medicamentosUsoHumano/informesPublicos/docs/2021/IPT_32-2021-Talzenna2.pdf

American Cancer Society. (s. f.). *Terapia hormonal para el cáncer de seno*. https://www.cancer.org/es/cancer/tipos/cancer-de-seno/tratamiento/terapia-hormonal-para-el-cancer-de-seno.html

Antena 3. (2024, 2 de octubre). *El precio de querer vivir: "Pagaré 11.500 euros cada tres semanas por mi tratamiento para el cáncer"*. https://www.antena3.com/programas/espejo-publico/noticias/precio-querer-vivir-pagare-11500-euros-cada-tres-semanas-tratamiento-cancer_2024100266fd30c9b3741e0001fcc201.html

Arrospide, A., Mar, J., Soto-Gordoa, M., Acaiturri, L., Basterretxea, M. J., Ezkurra, M., López-Linares, C., Linacisoro, I., & Idigoras, I. (2017). Comparación directa de los costes sanitarios en los dos últimos meses de vida de pacientes con cáncer. *Medicina Paliativa*, *24*(3), 154–161. https://doi.org/10.1016/j.medipa.2016.09.002

Asociación Española Contra el Cáncer. (s. f.-a). *1.100 € es el coste medio para un paciente que necesite recibir radioterapia*. https://www.contraelcancer.es/es/actualidad/noticias/1100eu-es-coste-medio-paciente-que-necesite-recibir-radioterapia

Asociación Española Contra el Cáncer. (s. f.-b). *Cáncer de mama*. https://www.contraelcancer.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama

Asociación Española Contra el Cáncer. (s. f.-c). *El impacto económico del cáncer en las familias en España*. Observatorio del Cáncer. https://observatorio.contraelcancer.es/informes/el-impacto-economico-del-cancer-en-las-familias-en-espana

Asociación Española Contra el Cáncer. (s. f.-d). *La radioterapia en España: costes asociados y ayudas a tener en cuenta*. Blog Contraelcancer.es. https://blog.contraelcancer.es/radioterapia-ayudas/

Asociación Española Contra el Cáncer. (s. f.-e). *Toxicidad financiera del cáncer de mama*. Observatorio del Cáncer. https://observatorio.contraelcancer.es/informes/toxicidad-financiera-del-cancer-de-mama

Beauty&Beauty. (s. f.). *El mejor precio de reconstrucción mamaria garantizado*. https://cirugiaesteticabyb.com/services/reconstruccion-mamaria/

Biotherapy International. (s. f.). *¿Cuánto cuesta la inmunoterapia contra el cáncer?* https://ibiotherapy.com/es/blog/cuanto-cuesta-la-inmunoterapia-contra-el-cancer/

Bookimed. (2024). *Terapia dirigida | TOP 10+ clínicas 2024 – Precio, médicos, reseñas*. https://es.bookimed.com/clinics/procedure=targeted-therapy/

Bookimed. (2025a). *Biopsia en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=biopsy/

Bookimed. (2025b). *Consulta con un oncólogo en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=consultation-of-oncologist/

Bookimed. (2025c). *Inmunoterapia en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=immunotherapy/

Bookimed. (2025d). *Mastectomía en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=mastectomy/

Bookimed. (2025e). *Quimioterapia en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=chemotherapy/

Bookimed. (2025f). *Radioterapia en España | Precio: TOP 10+ clínicas 2025*. https://es.bookimed.com/clinics/country=spain/procedure=radiotherapy/

Boutique Cáncer de Mama. (s. f.). *Prestaciones y ayudas*. https://www.boutique-cancer-de-mama.es/prestaciones-ayudas

Cadena SER. (2024, 23 de julio). *Las pacientes de cáncer de mama metastásico piden la financiación pública de dos fármacos*. https://cadenaser.com/euskadi/2024/07/23/las-pacientes-de-cancer-de-mama-metastasico-piden-la-financiacion-publica-de-dos-farmacos-radio-bilbao/

Casal-Mouriño, C., Varela-Lema, P., Lojo-Paz, S., Abad-López, L., Rúas-Álvarez, M., & Soto-García, M. J. (2021). Costes asociados al diagnóstico y tratamiento quirúrgico del cáncer de mama precoz. *Revista de Senología y Patología Mamaria*, *34*(2), 60–70. https://doi.org/10.1016/j.senol.2021.01.002

Clínica Magnasalud. (s. f.). *Ecografía mamaria*. https://clinicamagnasalud.es/servicios/ecografias/ecografia-mamaria/

Clínica Universidad de Navarra. (s. f.). *Segunda opinión a distancia*. https://www.cun.es/consulta-segunda-opinion

Comisión Interministerial de Precios de los Medicamentos. (2024). *Acuerdos de la sesión 247* [PDF]. Ministerio de Sanidad. https://www.sanidad.gob.es/areas/farmacia/precios/comisionInteministerial/acuerdosNotasInformativas/docs/ACUERDOS_CIPM_247.pdf

Comunidad de Madrid. (s. f.). *Evaluación económica de las pruebas genéticas en el tratamiento del cáncer de mama* [PDF]. https://www.comunidad.madrid/sites/default/files/aud/sanidad/evaluacion_economica_de_las_pruebas_geneticas_en_el_tratamiento_del_cancer_de_mama.pdf

Departamento de Sanidad de Aragón. (s. f.). *Costes medios sanitarios*. SaludInforma. https://www.saludinforma.es/portalsi/calidad-informacion-sanitaria/costes-medios-sanitarios

Dirección General de la Agencia de Calidad del SNS. (s. f.). *Boletín Impacto (Suplemento 26)*. https://www.sanidad.gob.es/organizacion/sns/planCalidadSNS/boletinAgencia/suplementoImpacto/26/actualidad1.html

e-Quirónsalud. (s. f.). *Panel de cáncer hereditario de mama y ovario*. https://e-quironsalud.es/estudios-geneticos/panel-de-cancer-hereditario-de-mama-y-ovario

elDiario.es. (2023, 10 de agosto). *Osakidetza revela que el coste medio de atención de un cáncer de mama en Euskadi es de 17.980 euros*. https://www.eldiario.es/euskadi/osakidetza-revela-coste-medio-atencion-cancer-mama-euskadi-17-980-euros_1_10855558.html

Espinàs, J. A., Borràs, J. M., Font, R., Torà, N., Solà, J., Quintana, M. J., & Galceran, J. (2013). Costes de los servicios sanitarios asociados al tratamiento del cáncer de mama precoz con radioterapia adyuvante. *Revista de Senología y Patología Mamaria*, *26*(3), 89–95. https://doi.org/10.1016/j.senol.2013.06.002

Estudio de minimización de costes en el tratamiento del cáncer de mama HER2+: Trastuzumab intravenoso vs subcutáneo. (s. f.). ILAPHAR – Revista de la OFIL. https://www.ilaphar.org/estudio-de-minimizacion-de-costes-en-el-tratamiento-del-cancer-de-mama-her2-trastuzumab-iv-vs-sc/

Farmacosalud. (2024). *Pacientes españolas reclaman la financiación y el acceso rápido a 5 fármacos para el cáncer de mama metastásico*. https://farmacosalud.com/pacientes-

espanolas-reclaman-la-financiacion-y-el-acceso-rapido-a-5-farmacos-para-el-cancer-de-mama-metastasico/

Forner-Cordero, I. (2016). Gestión económica del tratamiento del linfedema. *Revista de Senología y Patología Mamaria*, *29*(1), 1–2. https://doi.org/10.1016/j.senol.2016.02.001

Grupo de Trabajo de Farmacia Oncológica de la SEFH. (2024). Abemaciclib en adyuvancia para el tratamiento del cáncer de mama precoz con alto riesgo de recidiva. *Farmacia Hospitalaria*, *48*(2), 85–87. https://doi.org/10.1016/j.farma.2023.11.002

Grupo VIVO. (s. f.). *Resonancia magnética de mama*. https://grupovivo.life/producto/resonancia-magnetica-de-mama/

ILAPHAR – Revista de la OFIL. (s. f.). *Efectividad, seguridad y coste de la terapia oncológica en pacientes con cáncer de mama metastásico en la práctica clínica*. https://www.ilaphar.org/efectividad-seguridad-y-coste-de-la-terapia-oncológica-en-pacientes-con-cancer-de-mama-metastasico-en-la-practica-clinica/

Infobae. (2024, 16 de junio). *Estas son las ayudas que pueden solicitar los pacientes con cáncer en España*. https://www.infobae.com/espana/2024/06/16/estas-son-las-ayudas-que-pueden-solicitar-los-pacientes-con-cancer-en-espana/

Infobae. (2024, 27 de septiembre). *Sanidad financiará los medicamentos para las pacientes de cáncer de mama metastásico*. https://www.infobae.com/espana/2024/09/27/sanidad-financiara-los-medicamentos-para-las-pacientes-de-cancer-de-mama-metastasico/

International Agency for Research on Cancer. (s. f.). *Cancer Today*. https://gco.iarc.fr/today/en/data-sources-methods

La Nueva Crónica. (2024, 11 de enero). *El coste medio para un paciente que necesita radioterapia es de 1.100 euros*. https://www.lanuevacronica.com/actualidad/el-coste-medio-para-un-paciente-que-necesita-radioterapia-es-de-1-100-euros_89495_102.html

Mar, J., Arrospide, A., Barrutia, J., & Grupo Cáncer de Mama Euskadi. (2015). Coste del tratamiento del cáncer de mama por estadío clínico en el País Vasco. *Gaceta Sanitaria*, *29*(1), 23–28. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57272015000100010

Más Datos Cáncer. (2025, 15 de enero). *Más Datos Cáncer*. https://www.masdatoscancer.es/

Mi Diagnóstico. (s. f.-a). *Biopsia*. https://midiagnostico.es/categoria-producto/biopsia/

Mi Diagnóstico. (s. f.-b). *¿Cuál es el precio de una mamografía?* https://midiagnostico.es/cual-es-el-precio-de-una-mamografia/

Mi Diagnóstico. (s. f.-c). *Tu mamografía en Sevilla, desde 45 €*. https://midiagnostico.es/landing/mamografia-en-sevilla/

Ministerio de Sanidad. (2002). *Resolución de 26 de diciembre de 2001 sobre revisión de precios a aplicar por los centros sanitarios a las asistencias prestadas a terceros* [BOE-A-2002-215]. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2002-215

Ministerio de Sanidad. (2023a). *Informe de la situación sobre los pacientes largos supervivientes de cáncer en el Sistema Nacional de Salud*. https://www.sanidad.gob.es/areas/calidadAsistencial/estrategias/cancer/docs/Supervivientes_Cancer_ACCESIBLE.pdf

Ministerio de Sanidad. (2023b). *Registro de Actividad de Atención Sanitaria Especializada (RAE-CMBD): Actividad y resultados de la hospitalización en el SNS. Año 2022*. https://www.sanidad.gob.es/estadEstudios/estadisticas/docs/RAE-CMBD_Informe_Hospitalizacion_2022.pdf

Ministerio de Sanidad. (2023c). *Resolución 420/38235/2023, de 6 de junio, por la que se publica la Adenda al Convenio con la Comunidad de Madrid* [BOE-A-2023-14713]. https://www.boe.es/diario_boe/txt.php?id=BOE-A-2023-14713

Ministerio de Sanidad. (2025a). *Estadística de gasto sanitario público 2025*. https://www.sanidad.gob.es/estadEstudios/estadisticas/docs/EGSP2008/egspPrincipalesResultados.pdf

Ministerio de Sanidad. (2025b). *Perfiles nacionales de cáncer 2025: España* [Informe]. https://www.sanidad.gob.es/areas/calidadAsistencial/estrategias/cancer/docs/PERFILES_NACIONALES_DE_CANCER_2025_ESPANA.pdf

Ministerio de Sanidad. (s. f.). *Estudio del coste-efectividad de un programa de detección precoz del cáncer de mama en Cataluña* [Informe]. https://www.sanidad.gob.es/biblioPublic/publicaciones/recursos_propios/resp/revista_cdrom/VOL70/70_1_015.pdf

Ministerio de Sanidad. (s. f.). *Principales datos del sistema nacional de salud* [Informe]. https://www.sanidad.gob.es/estadEstudios/portada/docs/DATOS_SNS_02_2024.pdf

MPOIS. (2024, 10 de julio). *Valor de inmunoterapia: Beneficios, costos y opiniones*. https://mpois.com/2024/07/10/valor-de-la-inmunoterapia-beneficios-costos-y-opiniones/

Muface. (s. f.). *Ayudas para personas con enfermedad oncológica*. https://www.muface.es/muface_Home/Prestaciones/ayudas-proteccion-sociosanitaria/Ayudas-enfermos-oncologicos.html

Multiestetica. (s. f.). *Precio de reconstrucción mamaria*. https://www.multiestetica.com/precios/reconstruccion-mamaria

Oliver Wyman. (2018). *El impacto económico y social del cáncer en España* [Informe]. https://www.oliverwyman.es/content/dam/oliver-wyman/Iberia/Publications/el-impacto-economico-y-social-del-cancer-en-espana.pdf

Oliver Wyman. (2020, febrero). *El cáncer le cuesta a España al menos 19.300 millones de euros*. https://www.oliverwyman.es/es/media-center/2020/feb/el-cancer-le-cuesta-a-espana-al-menos-19-300-millones-de-euros.html

Omakase Consulting. (2018). *La carga del cáncer en España* [Informe]. https://www.omakaseconsulting.com/wp-content/uploads/2018/04/omakase-lab-3-2018--burden-of-cancer-in-spain.pdf

Operarme.es. (s. f.). *Resonancia magnética de mama*. https://www.operarme.es/resonancias-magneticas/precio-resonancia-magnetica-mama/

Organisation for Economic Co-operation and Development. (2024). *Abordando el impacto del cáncer en la salud, la economía y la sociedad: España*. https://www.oecd.org/es/publications/abordando-el-impacto-del-cancer-en-la-salud-la-economia-y-la-sociedad_ca1d8757-es/espana_48695950-es.html

Pérez-Morales, R., Amézcua-Gutiérrez, M. Á., Bargalló-Rocha, J. E., Cantellano-Orozco, M., Cortés-García, A., & Henne-Hernández, O. (2017). Análisis del costo-utilidad de los nuevos fármacos para tratamiento del cáncer de próstata metastásico resistente a la castración. *Revista Mexicana de Urología*, *77*(1), 26–34. https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-40852017000100026

PGS Consulting. (s. f.). *El costo económico del cáncer afecta la recuperación de los pacientes*. https://pgs-consulting.com/como-costear-tratamiento-de-cancer

PREVERAS. (s. f.). *Cáncer de mama y trabajo* [Documento]. https://www.preveras.org/docs/documentos/cancer_de_mama_y_trabajo.pdf

Redacción Médica. (2024, 17 de marzo). *La factura "invisible" del cáncer de mama ronda 18.000 euros por paciente*. https://www.redaccionmedica.com/autonomias/pais-vasco/la-factura-invisible-del-cancer-de-mama-ronda-18-000-euros-por-paciente-7969

Servicio Andaluz de Salud. (2023). *Aprepitant* [Ficha de evaluación]. https://www.sspa.juntadeandalucia.es/servicioandaluzdesalud/sites/default/files/sincfiles/wsas-media-mediafile_sasdocumento/2023/aprepitant.pdf

Servicio de Evaluación del Servicio Canario de la Salud. (s. f.). *Análisis coste-efectividad del cribado del cáncer de mama*. https://sescs.es/analisis-coste-efectividad-del-cribado-del-cancer-de-mama/

SmartSalus. (s. f.). *PET en Barcelona – Centros médicos y clínicas privadas al mejor precio*. https://www.smartsalus.com/barcelona/medicina-nuclear/pet

Sociedad Española de Farmacia Hospitalaria. (s. f.). *Estudio transversal del tratamiento del cáncer de mama en España* [PDF]. https://www.sefh.es/fh/92_2.pdf

Sociedad Española de Oncología Médica. (s. f.). *El cáncer en cifras*. https://seom.org/prensa/el-cancer-en-cifras

Sociedad Española de Senología y Patología Mamaria. (2023). *Consenso de seguimiento de pacientes con cáncer de mama* [PDF]. https://sespm.es/wp-content/uploads/2023/03/CONSENSO-DE-SEGUIMIENTO-DE-PACIENTES-CON-CANCER-DE-MAMA.pdf

Sociedad Española de Senología y Patología Mamaria. (2024a, 2 de octubre). *La Comisión de Precios rectifica y aprueba la financiación de dos fármacos revolucionarios para su indicación en cáncer de mama metastásico*. https://sespm.es/la-comision-de-precios-rectifica-y-aprueba-la-financiacion-de-dos-farmacos-revolucionarios-para-su-indicacion-en-cancer-de-mama-metastasico/

Sociedad Española de Senología y Patología Mamaria. (2024b, 24 de julio). *Sanidad niega la financiación de un fármaco revolucionario frente al cáncer de mama*. https://sespm.es/sanidad-niega-la-financiacion-de-un-farmaco-revolucionario-frente-al-cancer-de-mama/

TuMédico. (s. f.-a). *Desde 33 € – Ecografía mamaria en Madrid (reserva online)*. https://www.tumedico.es/radiologia-y-ecografia/ecografia-mamaria/madrid

TuMédico. (s. f.-b). *Desde 199 € – Resonancia magnética de mama*. https://www.tumedico.es/resonancia-magnetica/resonancia-magnetica-de-mama

TuMédico. (s. f.-c). *Desde 949 € – PET corporal*. https://www.tumedico.es/medicina-nuclear/pet

Vivolabs. (s. f.). *Estudio genético completo de cáncer hereditario*. https://vivolabs.es/producto/estudio-genetico-completo-de-cancer-hereditario/

9. Índices de figuras

9.1 Índice de tablas

9.2 Índice de ilustraciones

10. Anexos

10.1 Anexo 1: Código Python utilizado para el PCA de clasificación

```python
# PCA
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
# Escalar los datos
scaler = StandardScaler()
# Aplicar PCA
pca = PCA(n_components=2)
cancer_ampliado_pca = pca.fit_transform(cancer_ampliado.select_dtypes(include=[np.number]).drop(columns=['Diagnosis Result']))
# Convertir a DataFrame
cancer_ampliado_pca_df = pd.DataFrame(data=cancer_ampliado_pca, columns=['PC1', 'PC2'])
# Añadir la columna de Diagnosis Result
cancer_ampliado_pca_df['Diagnosis Result'] = cancer_ampliado['Diagnosis Result'].values
# Visualizar los resultados de PCA
plt.figure(figsize=(10, 6))
sns.scatterplot(data=cancer_ampliado_pca_df, x='PC1', y='PC2', hue='Diagnosis Result', alpha=0.7)
plt.title('PCA de Breast Cancer Dataset')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend(title='Diagnosis Result')
plt.show()

#varianza explicada por cada componente
explained_variance = pca.explained_variance_ratio_
print(f"Varianza explicada por cada componente: {explained_variance}")

#que variables incluye la componente 1 y 2
components = pca.components_
components_df = pd.DataFrame(components, columns=cancer_ampliado.select_dtypes(include=[np.number]).drop(columns=['Diagnosis Result']).columns, index=['PC1', 'PC2'])
print("Componentes de PCA:")
print(components_df)
```

10.2 Anexo 2: Código Python utilizado para el entrenamiento de modelos de clasificación

```python
#Definimos los modelos que queremos entrenar

from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC

modelos = {
    'Regresión Logística': LogisticRegression(max_iter=1000),
    'KNN (k=5)': KNeighborsClassifier(n_neighbors=5),
    'Árbol de Decisión': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42),
    'Gradient Boosting': GradientBoostingClassifier(random_state=42),
    'SVM (RBF)': SVC(kernel='rbf', probability=True, random_state=42),
}
```

10.3 Anexo 3: Código Python utilizado para la evaluación de modelos de clasificación

```python
#Definimos las métricas que queremos utilizar para evaluar el rendimiento del modelo para posteriormente comparar
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix


def evaluar_modelo(modelo, X_tr, X_te, y_tr, y_te):
    modelo.fit(X_tr, y_tr)
    y_pred = modelo.predict(X_te)
    y_proba = modelo.predict_proba(X_te)[:,1]

    # Extraer la matriz de confusión
    tn, fp, fn, tp = confusion_matrix(y_te, y_pred).ravel()

    return {
        'Accuracy':         accuracy_score(y_te, y_pred),
        'Precision':        precision_score(y_te, y_pred),
        'Recall':           recall_score(y_te, y_pred),
        'F1-score':         f1_score(y_te, y_pred),
        'ROC AUC':          roc_auc_score(y_te, y_proba),
        'Confusion Table': {
            'TP': tp,
            'TN': tn,
            'FP': fp,
            'FN': fn
        }
    }
```

10.4 Anexo 4: Código Python utilizado para el entrenamiento de modelos de regresión

```python
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from sklearn.model_selection import train_test_split
import numpy as np
import pandas as pd

# Regresión Lineal Múltiple
lin_reg_multiple = LinearRegression()
lin_reg_multiple.fit(X_train, y_train)
y_pred_multiple = lin_reg_multiple.predict(X_test)
```

```python
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# Modelo Ridge
ridge_model = Ridge(alpha=1.0)
ridge_model.fit(Xc_train, yc_train)
yc_pred_ridge = ridge_model.predict(Xc_test)
```

```python
from sklearn.ensemble import RandomForestRegressor

# Modelo Random Forest
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(Xc_train, yc_train)
yc_pred_rf = rf_model.predict(Xc_test)
```

10.5 Anexo 5: Código Python utilizado para la evaluación de modelos de regresión

```python
results = {
    'Modelo': ['Regresión Múltiple', 'Perceptrón (clasificación)'],
    'RMSE': [
        np.sqrt(mean_squared_error(y_test, y_pred_multiple)),
        np.nan  # No aplica a clasificación
    ],
    'R2': [
        r2_score(y_test, y_pred_multiple),
        np.nan
    ],
    'Accuracy (Perceptrón)': [
        np.nan,
        (y_pred_perceptron == (y_test > y_train.median()).values).mean()
    ]
}


results_df = pd.DataFrame(results)
print(results_df)
```

10.6 Anexo 6: Dashboard interactivo desarrollado con PowerBI

https://app.powerbi.com/reportEmbed?reportId=357a3ada-ce8c-44db-b7f3-5c06ba2efc05&autoAuth=true&ctid=032115c7-35fe-4637-b2c3-d0a42906ba7b