DESAFÍO LATAM

ACADEMIA DE TALENTOS DIGITALES

Predicción de estadio del cáncer de mama usando herramientas de Machine Learning

Proyecto Final Data Science G56

Profesor Guía:
Camilo González

Alumnos:
Macarena Araneda
Marlene Concha
Francisca Gálvez
Daniel Herrera
Jairo Rojas
Javier Rojas

Santiago, Julio 2023

Resumen

En los últimos años el cáncer ha demostrado un aumento significativo en su incidencia, y esto se atribuye a diferentes razones, como lo son: el exposoma de la vida temprana, las condiciones y estilo de vida, la alimentación de la persona, el envejecimiento de la población entre otros. El Departamento de Estadísticas e Información de Salud (DEIS) reportó en el año 2022 un total de 136.643 muertes en Chile, de las cuales 28.453 decesos corresponden a neoplasias, siendo la segunda causa de muerte por enfermedades del país, esto ha llevado al cáncer de mama a ser la primera causa de muerte de mujeres en edad reproductiva.

Anualmente se implementan políticas públicas orientadas a mejorar y reforzar la disponibilidad de atención preventiva, esto implica facilitar el acceso a exámenes imagenológicos, exámenes de laboratorio, exámenes patológicos y otros recursos relevantes, siempre priorizando la atención de los pacientes.

Un problema que se ha presentado es la falta de organización y estructuración de los datos, la cual, no facilita el análisis para la toma de decisiones informadas, ni para la implementación de modelos predictivos en la priorización de pacientes. En este contexto, se pueden implementar diversas tecnologías que apoyen a la organización de los datos, como data science y Machine Learning, que puedan entregar una priorización oncológica en ámbitos como la prevención y la detección temprana del cáncer.

Para abordar esta problemática, tras haber revisado el estado actual y profundizar en el tema, hemos desarrollado un modelo predictivo dentro de un marco de trabajo integral. Este modelo se enfoca en el cáncer de mama como caso de estudio, centrándose en el análisis de datos clínicos. Nuestra contribución representa una detección temprana de los estadios del cáncer, lo que resulta en una herramienta valiosa para la priorización de pacientes y una significativa aportación en el ámbito clínico y de gestión.

Palabras claves: Cáncer de mama, modelo predictivo, data science, Machine Learning, priorización.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Abreviaciones

FALP	: Fundación Arturo López Pérez
INE	: Instituto Nacional de Estadísticas
T	: Tamaño tumor
N	: Tumor en ganglio linfático
M	: Metástasis
FONASA	: Fondo Nacional de Salud
DEIS	: Departamento de estadísticas e información en salud

Capítulo 1. Introducción

1.1. Introducción General

Durante los últimos 20 años, el cáncer se ha transformado en una las principales causas de mortalidad a nivel mundial. De acuerdo a lo reportado por GloboCan de International Agency for Research on Cáncer, durante el año 2020, se diagnosticaron un total de 19,29 millones de personas con cáncer en el mundo [1].

En Chile, el Departamento de Estadísticas e Información de Salud (DEIS) reportó en el año 2022 un total de 136.643 muertes en Chile, de las cuales 28.453 decesos corresponden a neoplasias, siendo la segunda causa de muerte por enfermedades del país [2].

En el país, el cáncer de mama es la primera causa de muerte de mujeres en edad reproductiva. Dado lo anterior, cada año se trabaja en políticas públicas que permitan mejorar y reforzar la oportunidad de una atención con un enfoque preventivo, donde se pueda acceder a exámenes imagenológicos, de laboratorio, patológicos, entre otros. Además, priorizar la atención con especialistas cuando corresponda.

La gran cantidad de datos que existen en el país, por lo general, no se encuentran tabulados, ni estructurados, es por ello, que se debe recalcar la importancia de mejorar la calidad de los datos en salud. Con una data limpia y estructurada se pueden analizar datos y tomar decisiones informadas. Además, se puede trabajar en modelos predictivos, que pueden ser una herramienta de apoyo a la hora de priorizar pacientes.

Dado lo anterior, es que el objetivo de este proyecto es trabajar en un modelo que permita predecir el estadio del cáncer de mama, a partir de datos patológicos y clínicos asociados al cáncer.

Con esto, se espera generar una herramienta que permita predecir qué tan agresivo es o será el cáncer de mama. Lo anterior, podría ayudar a los referentes clínicos a priorizar la atención de pacientes según la gravedad de su neoplasia.

1.2. Objetivos

1.2.1 Objetivo General

El objetivo general de este trabajo es desarrollar un modelo predictivo utilizando herramientas de Machine Learning para la detección temprana de los estadios del cáncer de mama según la gravedad de su neoplasia, con el fin de proporcionar una herramienta de priorización de pacientes y contribuir significativamente al ámbito clínico y de gestión.

1.2.2 Objetivos Específicos

- Realizar un análisis exploratorio de los datos recopilados para identificar patrones, tendencias y relaciones relevantes entre las variables clínicas y los estadios del cáncer de mama.
- Diseñar y desarrollar un modelo predictivo utilizando algoritmos de Machine Learning, como clasificación o regresión, para predecir los estadios del cáncer de mama en función de las variables clínicas disponibles.
- Evaluar y comparar el rendimiento de diferentes modelos de Machine Learning utilizando métricas adecuadas, como precisión, recall y área bajo la curva ROC, para determinar el modelo más efectivo en la predicción de los estadios del cáncer de mama.
- Implementar una interfaz o herramienta accesible que permita a los profesionales de la salud ingresar los datos clínicos de un paciente y obtener una predicción del estadio del cáncer de mama.
- Analizar el impacto potencial del modelo predictivo en la priorización de pacientes y su aplicación en la práctica clínica, considerando factores como la optimización de recursos, la toma de decisiones informadas y la mejora en la detección temprana del cáncer de mama.

1.3. Requerimientos

El objetivo es desarrollar un proyecto que priorice la atención médica a los pacientes con cáncer de mama, entregando un enfoque predictivo y basado en datos entregados por el equipo de Informática Médica y Data Science de la FALP (Fundación Arturo López Pérez). Para conseguirlo se han establecido una serie de requisitos que guiarán el desarrollo de un modelo de clasificación capaz de predecir el estadio del cáncer de mama a partir de datos patológicos y clínicos asociados.

Con la finalidad de cumplir con el objetivo planteado es que buscamos realizar los siguientes procedimientos:

- Mejorar la calidad de los datos en salud para el análisis y la toma de decisiones informadas. Se implementarán estrategias para potenciar cualitativamente los datos relacionados con el cáncer de mama, asegurando que estén tabulados y estructurados de manera adecuada. Esto permitirá realizar análisis más precisos y tomar decisiones basadas en datos confiables.
- Desarrollar modelos predictivos que permitan priorizar pacientes de riesgo. Se trabajará en generar predicciones que puedan ayudar en la decisión de priorizar a los pacientes en función de la agresividad del cáncer. Estos modelos se utilizarán como herramientas de apoyo para la toma de decisiones en los departamentos de salud y la asignación de recursos.
- Evaluar y optimizar el rendimiento del modelo de predicción. Se realizarán pruebas al modelo de predicción desarrollado, evaluando su precisión y rendimiento. Además, se realizaron ajustes en el modelo para optimizar su capacidad de predicción y garantizar resultados confiables.

El proyecto nos brindará una herramienta que permita predecir el estadio del cáncer de mama. Categorizando según ciertas características clínicas si el cáncer es Inicial – Intermedio – Avanzado. Esta herramienta, sería un apoyo a los gestores de casos oncológicos y a quienes deben priorizar a los pacientes.

1.4. Equipo de trabajo

Con el fin de ejecutar el proyecto actual, los miembros del equipo han establecido previamente responsabilidades que se centrarán en áreas específicas del trabajo, con el objetivo de mantener un orden durante el proceso laboral. Los roles fueron determinados considerando las habilidades y características individuales de los integrantes, y se detallan a continuación:

Líderes: Macarena Araneda - Daniel Herrera.

- Mantener al equipo actualizado sobre el progreso del proyecto.
- Establecer metas y objetivos claros para finalizar las entregas.
- Coordinar y delegar tareas a los miembros del equipo.
- Tomar decisiones estratégicas y resolver problemas que surjan durante el proyecto.

Analistas de datos: Marlene Concha - Jairo Rojas.

- Estudiar los datos disponibles y comprender su estructura y características.
- Realizar análisis exploratorio de los datos para identificar patrones, tendencias y relaciones.
- Preprocesar y limpiar los datos para su posterior análisis.
- Enriquecer la base de datos mediante técnicas de extracción, transformación y carga.
- Aplicar técnicas estadísticas y de minería de datos para obtener información relevante.

Control de calidad y validación de datos: Jairo Rojas - Francisca Gálvez.

- Garantizar la calidad de los datos existentes o agregados al proyecto.
- Realizar verificaciones y validaciones de los datos para asegurar su integridad y coherencia.
- Realizar procesos de limpieza, normalización y codificación de los datos.
- Implementar medidas para prevenir errores y detectar anomalías en los datos.

Ingenieros de modelamiento: Jairo Rojas - Javier Rojas - (Apoyo: Macarena Araneda - Daniel Herrera)

- Proponer y diseñar modelos predictivos o descriptivos basados en los datos disponibles.
- Implementar algoritmos y técnicas de Machine Learning o estadística para crear los modelos.
- Evaluar y comparar diferentes modelos para seleccionar el más adecuado.
- Ajustar y optimizar los modelos para obtener resultados precisos y confiables.
- Realizar pruebas y validaciones de los modelos para asegurar su rendimiento y eficacia.

Visualización de datos: Javier Rojas - Marlene Concha.

- Crear representaciones visuales de los datos para facilitar su comprensión y análisis.
- Diseñar gráficos, tablas y dashboards interactivos que muestren de manera clara y efectiva la información recopilada.
- Identificar las visualizaciones más adecuadas para comunicar los resultados del análisis de datos.
- Presentar los datos de manera atractiva y comprensible para los usuarios y stakeholders del proyecto.

Documentadores: Daniel Herrera - Macarena Araneda - Francisca Gálvez.

- Registrar y documentar los pasos, procesos y decisiones tomadas durante el proyecto.
- Mantener un registro actualizado de los cambios realizados en los archivos y documentos del proyecto.
- Crear documentación clara y completa que describa los métodos, técnicas y resultados obtenidos.
- Organizar y archivar las versiones anteriores de los documentos y archivos para facilitar el seguimiento y la referencia.

Capítulo 2. Planificación de la investigación

2.1. Definición de la pregunta de investigación

¿Es posible desarrollar un modelo de clasificación supervisado basado en variables patológicas, clínicas y sociodemográficas para predecir el estadio del cáncer de mama y priorizar la atención médica de los pacientes?

2.2. Hipótesis de trabajo

Se espera que, al utilizar datos patológicos y clínicos asociados al cáncer de mama, sea posible desarrollar un modelo de clasificación supervisado que permita acercarnos a la predicción del estadio del cáncer en pacientes. Este modelo ayudará a mejorar la priorización de la atención médica, optimizando la gestión de pacientes.

2.3. Definición del Vector Objetivo

El vector objetivo será la variable **STATUS** que determinaremos según la agrupación de la variable **ESTADIO.** El estadio de un cáncer de mama se determina en función de sus características, como su tamaño, y/o receptores hormonales definidos en **TNM** (más información en Anexos).

2.4. Estrategias analíticas a nivel descriptivo

Las estrategias analíticas a nivel descriptivo se enfocan en explorar, caracterizar y comprender datos relacionados con el cáncer de mama, con el objetivo de obtener una visión general de las variables y su distribución, identificar patrones y tendencias y establecer las bases para un mayor desarrollo del modelo de clasificación:

- **Análisis exploratorio de datos**: Se realizará un análisis exploratorio de datos relacionados con el cáncer de mama utilizando la base de datos entregada por FALP. Esto implica examinar la distribución de datos, identificar valores atípicos, analizar la correlación entre variables, explorar datos faltantes y hacer gráficos descriptivos para visualizar patrones y tendencias.

- **Caracterización de variables**: Realizar una descripción detallada de las variables utilizadas en el modelo, tanto patológicas como clínicas. Esto puede incluir el cálculo de medidas de resumen como la media, la mediana, la desviación estándar, el máximo y el mínimo para variables numéricas. Para variables categóricas, se pueden calcular frecuencias y porcentajes para identificar las categorías más comunes.

- **Análisis de distribución**: Evaluar la distribución de variables relacionadas con el cáncer de mama. Esto puede implicar dibujar histogramas, diagramas de caja y diagramas de densidad para comprender la forma y la dispersión de los datos. Además, se pueden utilizar pruebas estadísticas para determinar si las variables siguen una distribución normal o si requieren transformaciones.

- **Análisis de correlación**: Investigar las conexiones entre variables patológicas y clínicas. Esto se puede lograr calculando la matriz de correlación y visualizándola mediante un mapa de calor. Identificar las relaciones más fuertes y débiles entre las variables puede ayudar a comprender mejor cómo se relacionan y qué variables pueden ser más relevantes para predecir el estadio del cáncer de mama.

- **Segmentación de datos**: Los datos, se deben segmentar, según las características relevantes como la edad, el sexo u otras variables clínicas importantes. Esto permitirá analizar y comparar diferentes grupos de pacientes para identificar posibles diferencias en los estadios del cáncer de mama y brindar información adicional para el modelo de clasificación.

2.5. Consideraciones

Para el presente proyecto se deben tener las siguientes consideraciones:

- Es crucial garantizar la seguridad y privacidad de la información del paciente, por lo cual la base de datos carece de datos sensibles.
- La base de datos fue proporcionada por la Fundación Arturo López Pérez (FALP), la cual posee datos nulos, y se aplicarán técnicas para obtener mejores métricas.
- La base de datos corresponde a una muestra, que además se filtró por la categoría más frecuente, la cual es Cáncer de Mama.
- La modelación adecuada para el vector objetivo es de clasificación multiclase supervisado, donde se aplicará las técnicas de modelamiento y métricas correspondientes
- El dataset recibe el nombre de Registro_Tumores_FALP.csv, lo cual corresponde a la muestra total de datos.

2.6. Carta Gantt

TÍTULO DEL PROYECTO	Proyecto Data Science G56 - Cáncer de mama	**PROFESOR GUÍA**	Camilo González
RESPONSABLES DEL PROYECTO	Macarena Araneda - Marlene Concha - Francisca Gálvez - Daniel Herrera - Jairo Rojas - Javier Rojas	**FECHA**	12/06/23

NÚMERO EDT	TÍTULO DE LA TAREA	RESPONSABLE DE LA TAREA	FECHA DE INICIO	FECHA DE ENTREGA	HITO 1					HITO 2					HITO 3					HITO 4				
					SEMANA 1					SEMANA 2					SEMANA 3					SEMANA 4				
					5	6	7	8	9	12	13	14	15	16	19	20	21	22	23	26	27	28	29	30
1	**Definición y planificación del proyecto**																							
1.1	Análisis de la base de datos FALP	Todos	07/06/23	10/06/23																				
1.1.1	Reflexión y planificación del trabajo	Todos	05/06/23	07/06/23																				
1.2	Documentación Informe y Carta Gantt	Macarena Araneda y Daniel Herrera	10/06/23	29/06/23																				
1.3	Definición del alcance y de los objetivos	Todos	08/06/23	10/06/23																				
2	**Análisis Exploratorio de los datos**																							
2.1	Caracterización de variables	Todos	12/06/23	14/06/23																				
2.2	Segmentación de datos	Todos	13/06/23	15/06/23																				
3	**Preprocesamientos de los datos**																							
3.1	Limpieza, estructuración de los datos y Feature engineering.	Javier Rojas y Jairo Rojas	12/06/23	13/06/23																				
4	**Análisis y visualización de datos**																							
4.1	Análisis descriptivo y gráficos relevantes para la investigación	Marlene Concha	13/06/23	15/06/23																				
4.2	Análisis de distribución y de correlación de las variables	Marlene Concha	14/06/23	16/06/2023																				
4.3	Análisis de resultados	Todos	20/06/23	21/06/23																				
5	**Modelamiento**																							
5.1	Entrenamiento de modelos candidatos	Jaime Rojas y Jairo Rojas	12/06/23	16/06/23																				
5.2	Selección de modelos e hiperparámetros	Javier Rojas y Jairo Rojas	19/06/23	21/06/23																				
5.3	Re-evaluación de modelos	Javier Rojas y Jairo Rojas	21/06/23	23/06/23																				

Capítulo 3. Análisis exploratorio de datos Iniciales

El análisis de los datos entregados por el equipo de Informática Médica y Data Science de la FALP se acotó exclusivamente al cáncer de mama debido a su gran relevancia en relación con la cantidad de muertes anuales a nivel nacional. Para poder comprender de mejor manera cómo es que se distribuyen estos datos y de qué manera se relacionan es que se realiza un análisis exploratorio previo al preprocesamiento de los datos.

Cabe resaltar que los datos pertenecientes a **Distribución por descripción clínica** corresponden a exámenes que se ha realizado el paciente durante su diagnóstico por lo que es posible que una persona no se los haya realizado todos y tenga valores nulos en algunos de ellos.

También, que la mayor parte de la muestra se encuentra viva y reside en la región metropolitana, con un 7.26% de presencia en la comuna de “Las Condes”. Respecto a los tratamientos, la mayoría se ha realizado Quimioterapia, Cirugías, Radioterapia e Inmunoterapia (88.24%).

3.1 Análisis de valores perdidos

A continuación, se presentan gráficos de valores perdidos por filas en la matriz de valores nulos y un gráfico de barra con la cantidad de valores perdidos por campo.

Como se puede observar en la figura 1, el *df* cuenta con 5.575 filas con datos, 23 variables (que se encuentran definidas en el anexo A.2) de las cuales 3 contienen datos de tipo *int64* y 21 de tipo *object.*

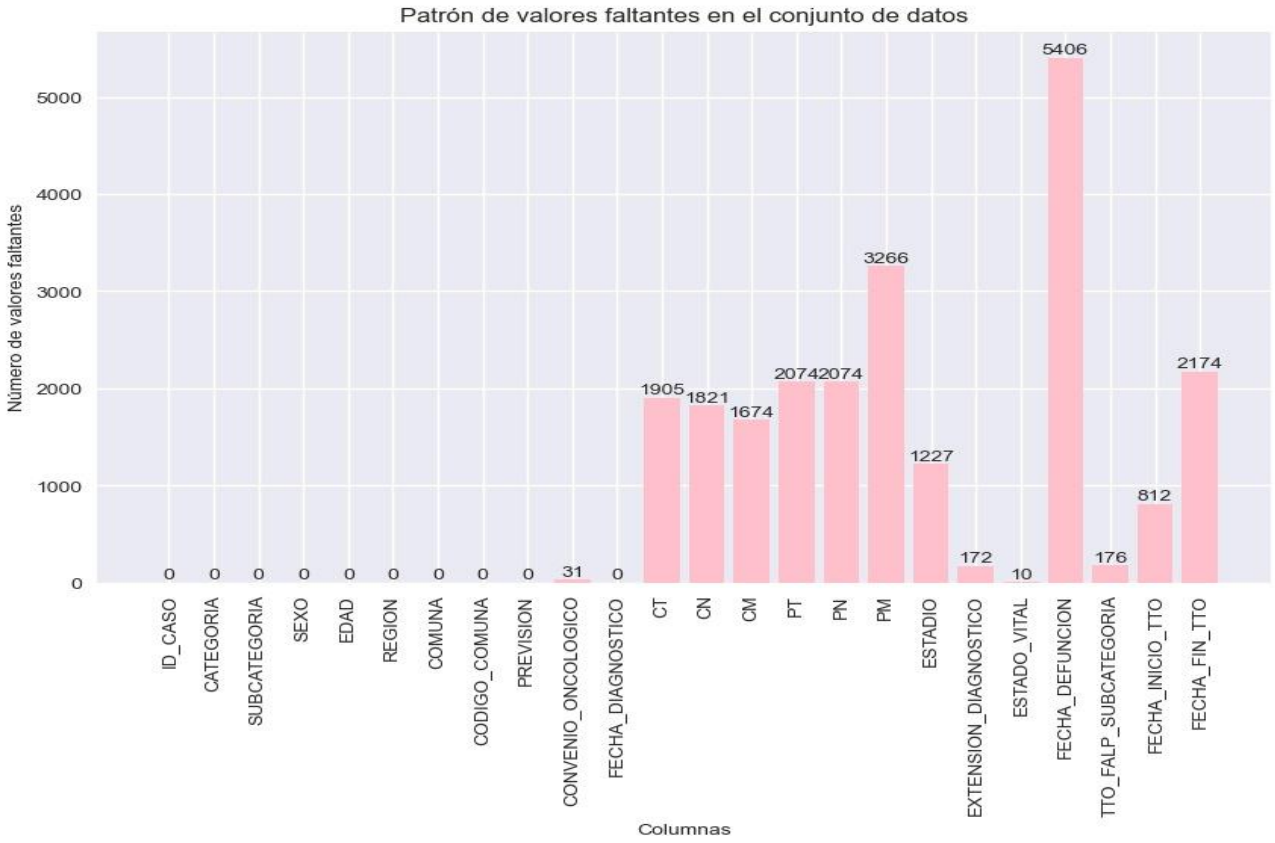

Figura 1 Valores perdidos

De estas 23 variables, 14 presentan valores *null*s: CONVENIO_ONCOLOGICO, CT, CN, CM, PT, PN, PM, ESTADIO, EXTENSION_DIAGNOSTICO, ESTADO_VITAL, FECHA_DEFUNCION, TTO_FALP_SUBCATEGORIA, FECHA_INICIO_TTO y FECHA_FIN_TTO.

3.2 Análisis Univariado

Variable sexo

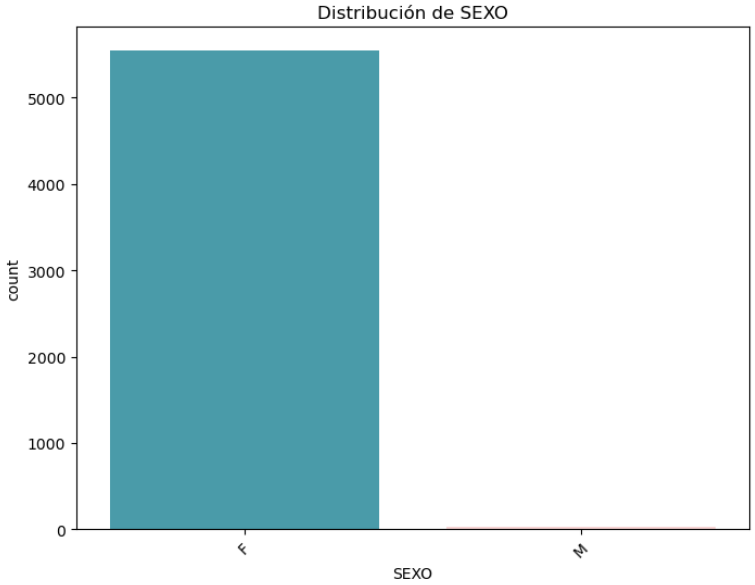

Figura 2Sexo F: Femenino M: Masculino

Respecto a la variable SEXO, se observa que la mayoría de los datos pertenecen al género femenino (5.542 vs 33), esto demuestra la tendencia de que este tipo de cáncer afecta mayoritariamente a las mujeres, sin embargo, no anula la posibilidad de que un hombre pueda padecer de esta enfermedad.

Variable extensión del diagnóstico

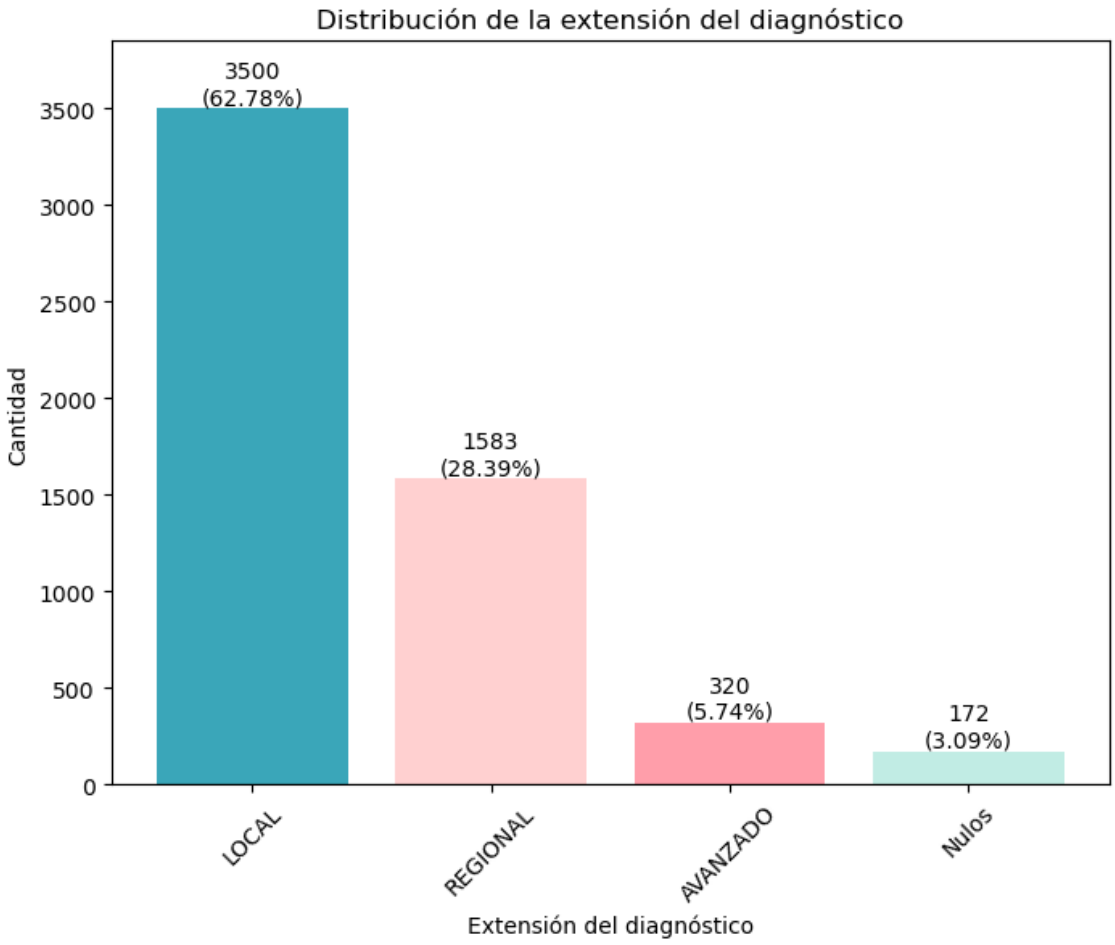

Figura 3 Distribución por tipo de extensión del cáncer

En la extensión del diagnóstico se evalúa el alcance del cáncer de mama, es decir, hasta qué partes del cuerpo se ha ido expandiendo. Como se puede observar en la figura 3, la mayoría se encuentra en la categorización “Local”, es decir, el cáncer se encuentra solamente en la mama. Como segunda opción más frecuente se observa que el cáncer ya ha comenzado a afectar a los ganglios linfáticos (“Regional”) y por último se encuentra el estado “Avanzado” en dónde se indica que el cáncer se encuentra en otras partes del cuerpo aparte de la mama.

Variable estadio

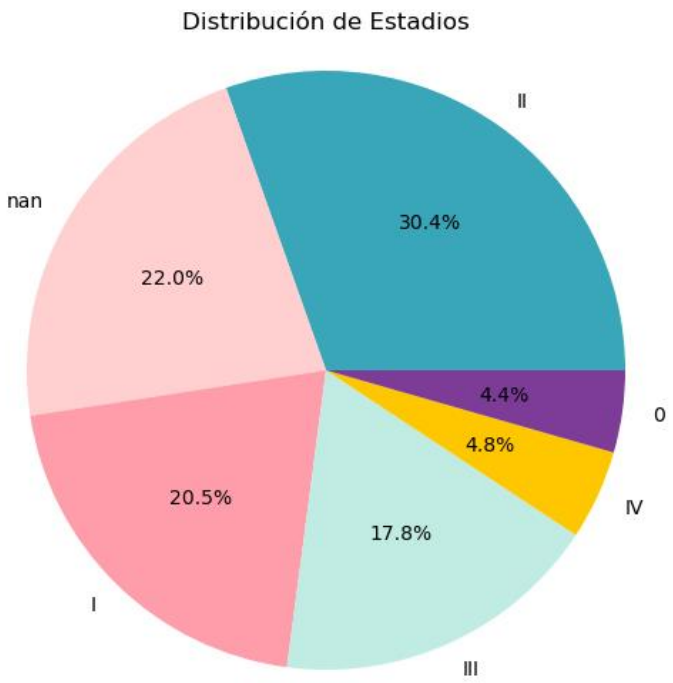

Figura 4 Distribución por tipo de Estadio del cáncer

En el caso de la variable ESTADIO, la cual corresponde a nuestro vector objetivo y nos indica la extensión y gravedad del cáncer, se observa que sus datos son de tipo *object* y que la mayor concentración de este se encuentra en el Estadio II con un 30.4%, seguido por el Estadio I con un 20.5%, el Estadio III con un 17.8%, el Estadio IV con sólo un 4.8% y por último el Estadio 0 correspondiente 4.4% presente en la muestra. También se logra observar una gran cantidad de valores faltantes (22.0%). El cáncer en estadio II nos indica que el tumor existente mide entre 2 a 5 cm y puede o no afectar a los ganglios linfáticos, mientras que el estadio I nos indica que el cáncer es pequeño y solamente se encuentra en la mama.

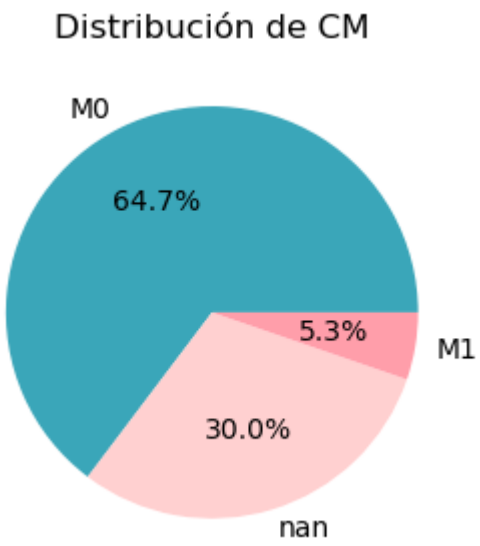

Figura 5 Distribución por descripción clínica de la metástasis

En la figura 5 se observa que un 64.7% de los pacientes tienen un resultado de M0, lo que nos indica que no hay evidencia de metástasis a distancia en el momento del diagnóstico, es decir que el cáncer sólo se encuentra en la mama y puede que en los ganglios linfáticos regionales. En el caso de M1, que es la minoría de la muestra, nos indica que el cáncer se ha diseminado a órganos distantes. Existe un 30% de pacientes que no registran valores en este examen

.

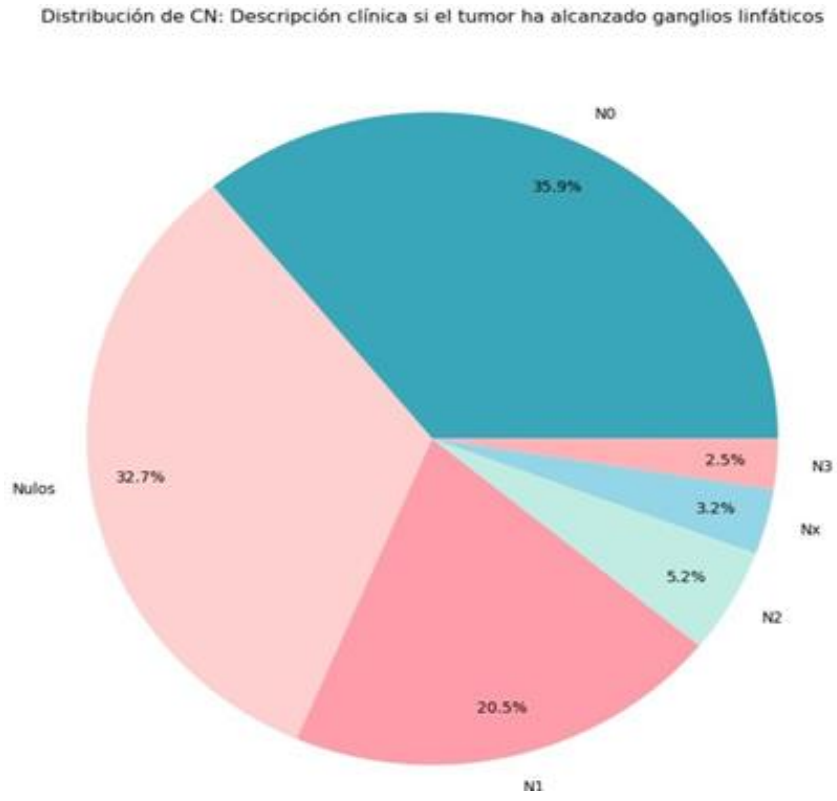

Figura 6 Distribución de la descripción clínica del alcance de ganglios linfáticos

En el caso de CN, nos indica la clasificación del estado de metástasis en los ganglios linfáticos regionales. En la data, se presenta N0 con un 35.9% lo que nos indica que no se encontraron ganglios linfáticos regionales afectados mientras que N1 se presenta con un 20.5% indicando la presencia de metástasis en los ganglios linfáticos axilares. También hay que destacar que un 32.7% de la muestra no registra valores en este procedimiento.

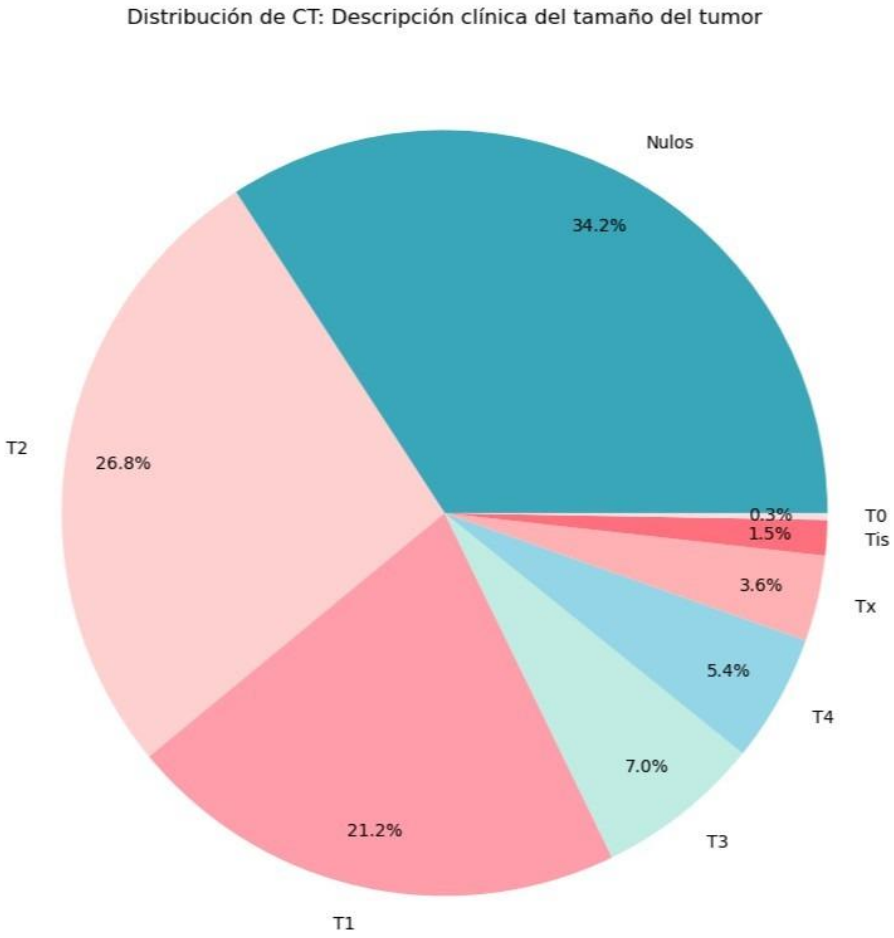

Figura 7 Distribución de la descripción clínica del tamaño del tumor

En el caso de la variable “CT”, esta nos indica la clasificación clínica del tamaño y la extensión del tumor primario (en base a exámenes clínicos e imágenes). La data presenta mayor predominancia en T2 (26.8%) lo que indica que el tumor mide entre 2 y 5 centímetros pudiendo o no existir afectación a estructuras adyacentes. Le sigue T1 factor que indica que el tumor mide solamente hasta 2 cm pudiendo o no existir afectación a estructuras adyacentes.

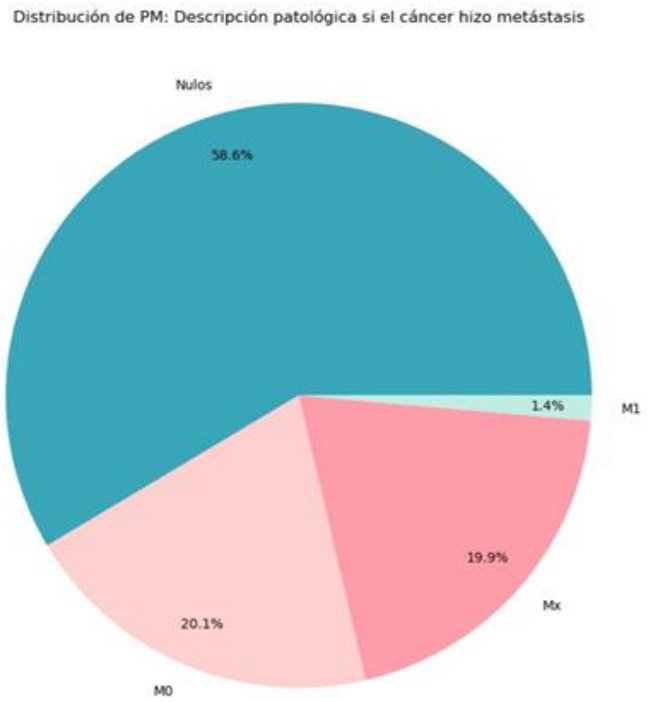

Figura 8 Distribución por descripción patológica de la metástasis

La variable “PM” indica la clasificación del estado de metástasis en los ganglios linfáticos mamarios internos. En la data se observa que la mayor parte de los pacientes (56.6%) no registra valores para esta variable, es decir, no se ha realizado la medición, y que la mayoría que sí presenta información, se encuentra en M0 con un 20.1% indicando que no hay presencia de células cancerosas en los ganglios. Le sigue Mx con un 19.9% que inca que esta información no fue evaluada o no está disponible y por último se encuentra M1 que indica que si existe evidencia de metástasis en los ganglios linfáticos mamarios internos, siendo esta la minoría en la data.

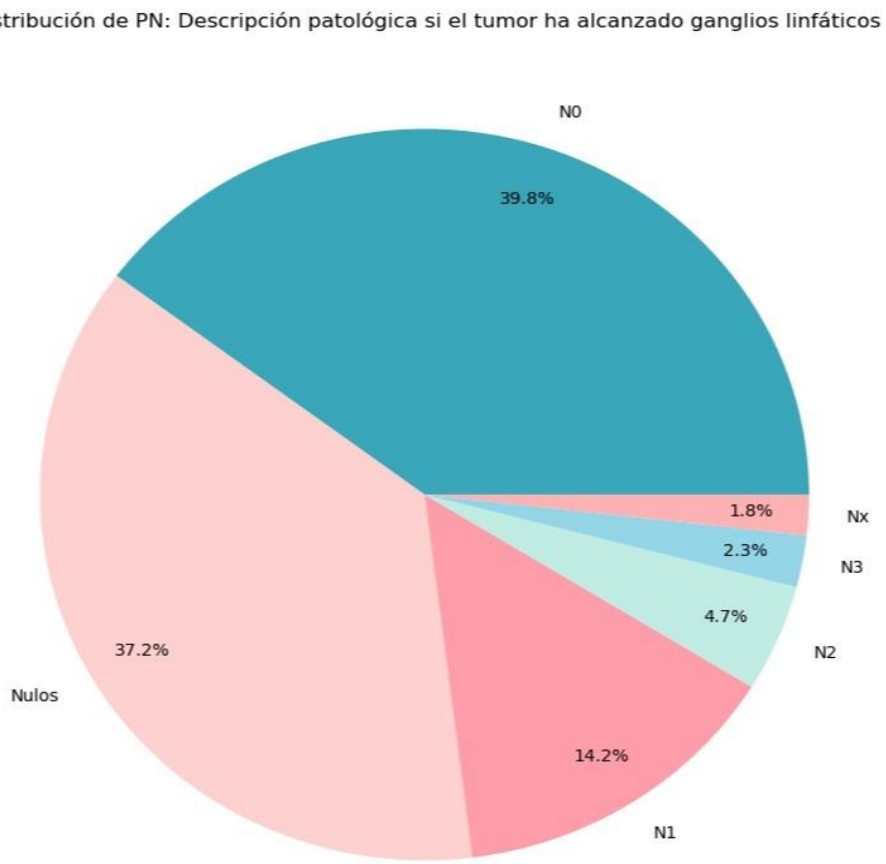

Figura 9 Distribución de la descripción patológica del alcance de ganglios linfáticos

En la figura 9 se puede observar la clasificación de la afectación de los ganglios linfáticos regionales más conocida como “PN”. La mayor parte de la muestra tiene una clasificación N0 que indica que no se encontró evidencia de metástasis en los ganglios linfáticos regionales. Seguida por N1 que indica presencia de metástasis en los ganglios linfáticos axilares del mismo lado del cuerpo que el tumor primario. Hay que destacar que existe un 37.2% de valores nulos para esta variable.

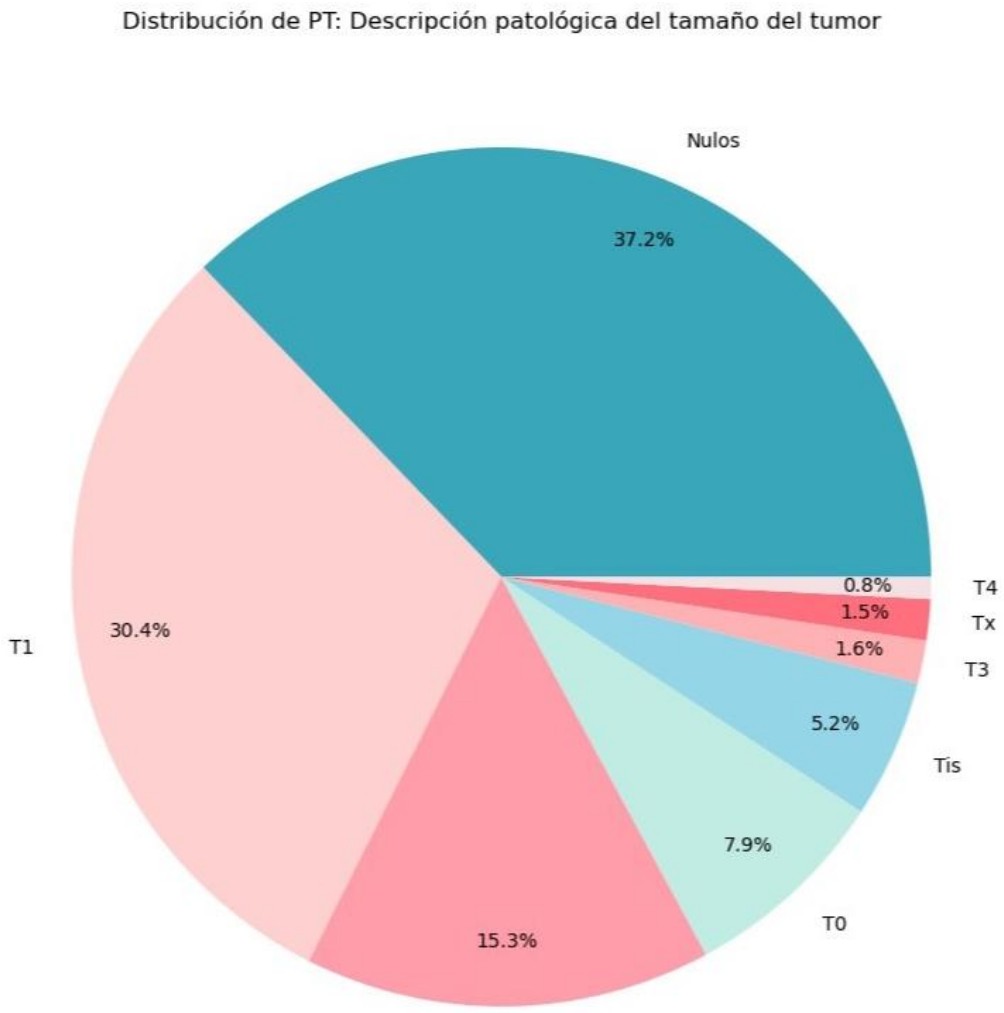

Figura 10 Distribución de la descripción patológica del tamaño del tumor

En el caso de la variable “PT”, esta nos indica la clasificación patológica del tamaño y la extensión del tumor primario (en base al análisis microscópico del tejido tumoral). La data presenta mayor predominancia en T1 (30.4%) lo que indica que el tumor mide hasta 2 centímetros y se encuentra confinado en la mama sin invadir estructuras adyacentes. Le sigue T2 con 15.3% factor que indica que el tumor mide entre 2 y 5 cm de diámetro y no ha invadido las estructuras adyacentes. Destacar también que existe un 37.2% de valores nulos para esta variable.

Variable convenio oncológico

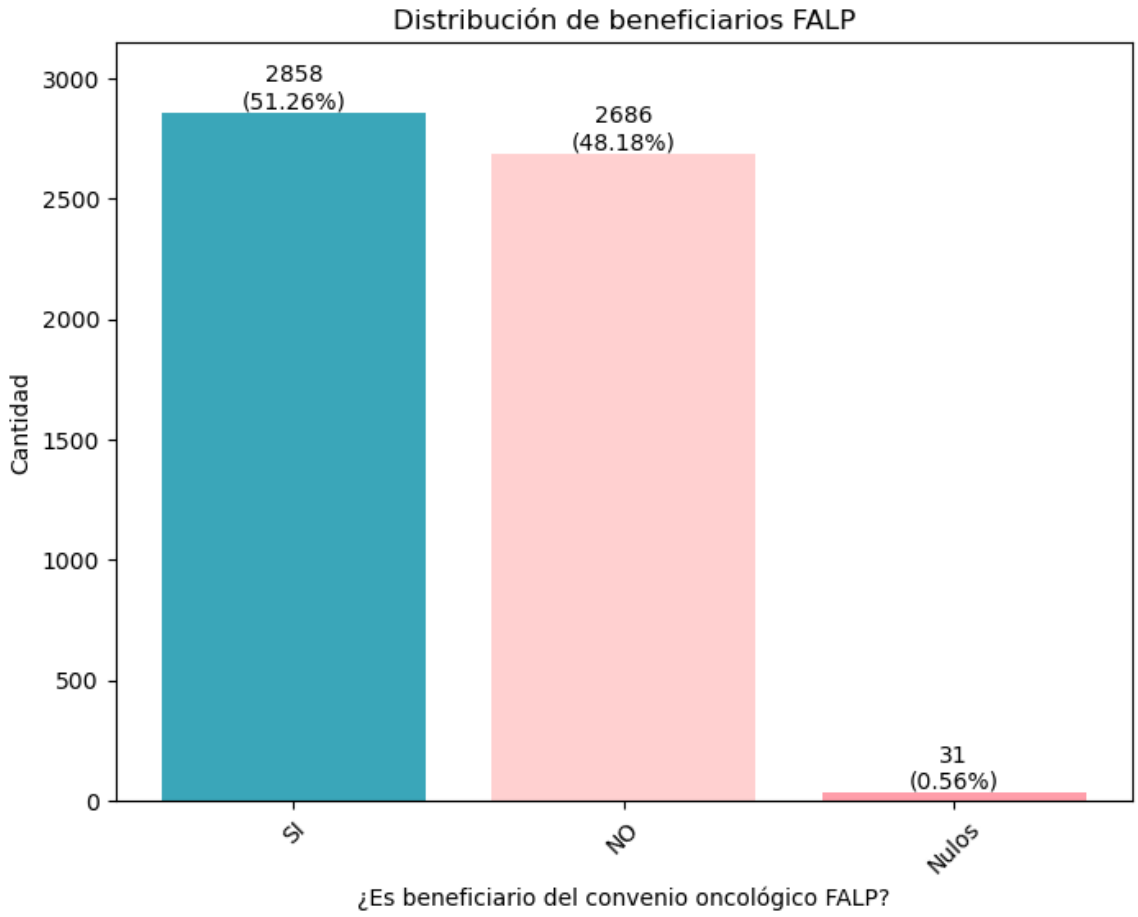

Figura 11 Distribución de convenio oncológico en los pacientes

En el gráfico de la figura 11, podemos observar que un 51.26% de los pacientes son beneficiarios de la FALP y que sólo un 0.56% no registra información en esta variable. Esta mayoría, se asocia a que la base de datos pertenece a la FALP, por lo que, la trazabilidad de pacientes de otros centros derivados acá, no se refleja en tiempo real en esta tabla de datos, su registro dependerá de diversos factores, entre ellos, si el paciente lleva a la FALP todo su expediente o no.

Variable previsión

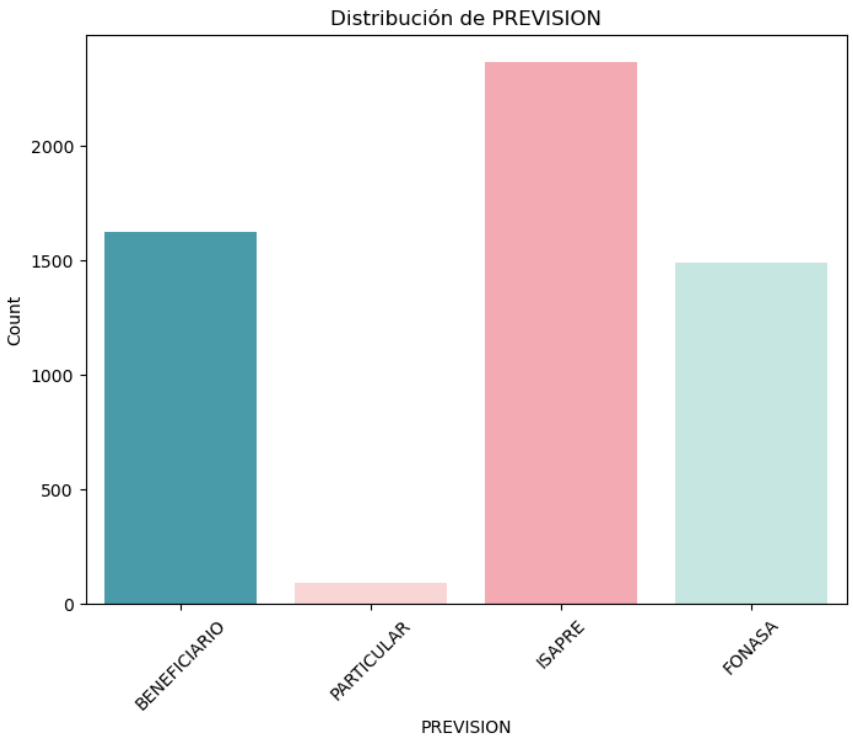

Figura 12 Distribución de la previsión que pertenecen los pacientes

Se observa que la mayor parte de los pacientes pertenecen al sistema de salud privado de Chile ISAPRE. La segunda mayoría corresponde a beneficiarios de convenio oncológico y la tercera mayoría a pacientes de FONASA. Esto se atribuye a que la FALP, tiene un modelo de convenio donde los pacientes de ISAPRE o FONASA que cuentan con un convenio oncológico pueden acceder a una atención como de la FALP como prestador principal de salud o como segundo prestador de salud, en caso de que en el centro de atención del paciente no puedan entregar el tratamiento en los tiempos necesarios (por urgencia, lista de espera, tiempos GES u otro).

Variable edad

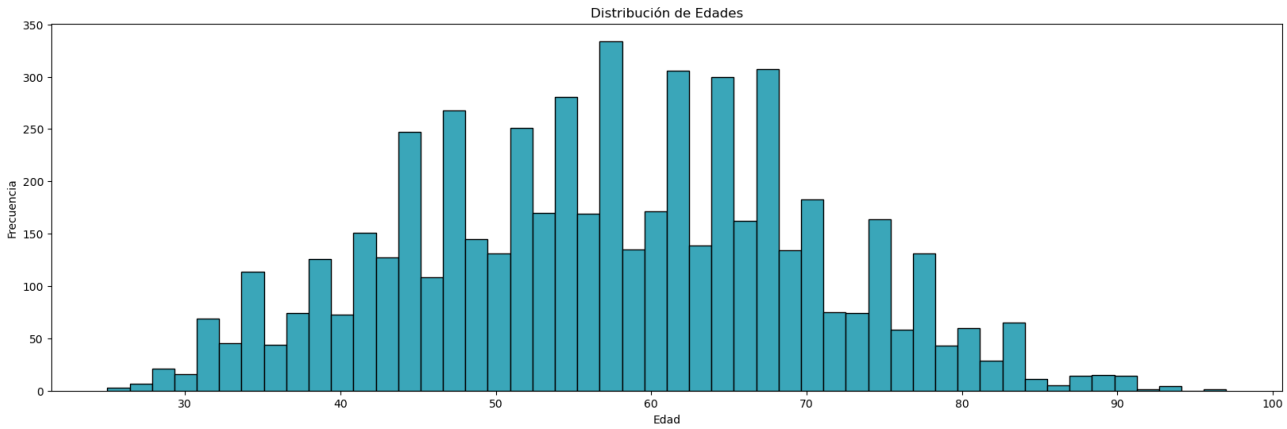

Figura 13 Distribución de las Edades

Respecto a la variable edad, se puede observar que no presenta valores nulos para ninguna fila, su moda es de 57 años existiendo una concentración entre los 48 y 66 años en la data. El individuo más joven para analizar tiene 25 años mientras que el más longevo tiene 97 años. Con el objetivo de reducir la granularidad de los datos, poder resumirlos de mejor forma y facilitar su comparación y análisis es que evaluará recodificar esta variable en "Grupos Etarios".

Capítulo 4. Preprocesamiento de los datos

Para el preprocesamiento de la base de datos se realizó lo siguiente:

- Se generó una copia de la base de datos, para no alterar los datos originales.
- Se filtraron los datos según la columna “CATEGORIA”, dejando únicamente las filas que tengan el valor “MAMA”.
- Se optó por recodificar todas las variables fechas (FECHA_DIAGNOSTICO, FECHA_DEFUNCION, FECHA_INICIO_TTO y FECHA_FIN_TTO) de un string a un datetime. Luego, se calculó la duración de los tratamientos (FECHA_FIN_TTO - FECHA_INICIO_TTO) y la espera entre el diagnóstico y el inicio del tratamiento (FECHA_INICIO_TTO - FECHA_DIAGNOSTICO). Las columnas calculadas fueron agregadas al dataframe.
- Para el modelamiento no se consideraron los datos de fechas.
- Se agrupó la columna EDAD según el rango etario utilizado por el INE, como se muestra en la Tabla 1: (Niñez (0-14 años), Jóvenes (15 - 29 años), Adultos (30-59) y vejez (14 - 26 años), adultez (27 - 59 años) y vejez (60 años y más)).

Tabla 1Rango etario

Grupos etarios	**Edad (años)**
Niños	0 – 14
Jóvenes	15 – 29
Adultos	30 – 59
Adulto mayor	≥ 60

- Se creó la columna STATUS (vector objetivo) utilizando la variable ESTADIO. Si la variable ESTADIO era 0 o I, entonces se asignó el valor “1” (Inicial). Si la variable ESTADIO era II o III, entonces se asignó el valor “2” (Intermedio). Finalmente, si la variable ESTADIO era IV, entonces se asignó el valor “3” (Avanzado).

Tabla 2 Variable STATUS

ESTADIO	Valor asignado STATUS	Significado
0	1	Inicial
I	1	Inicial
II	2	Intermedio
III	3	Avanzado
IV	3	Avanzado

- Todas las variables categóricas restantes fueron transformadas a variables indicadoras (dummies).
- Se eliminaron todas las filas que tuvieran un valor faltante en la variable ESTADIO.
- Se reemplazaron los valores NaN con el valor “No Determinado” en aquellas filas que al menos tengan un valor distinto a NaN en las columnas que definen el detalle del tumor (TC, CN, CM, PT, PN, PM).

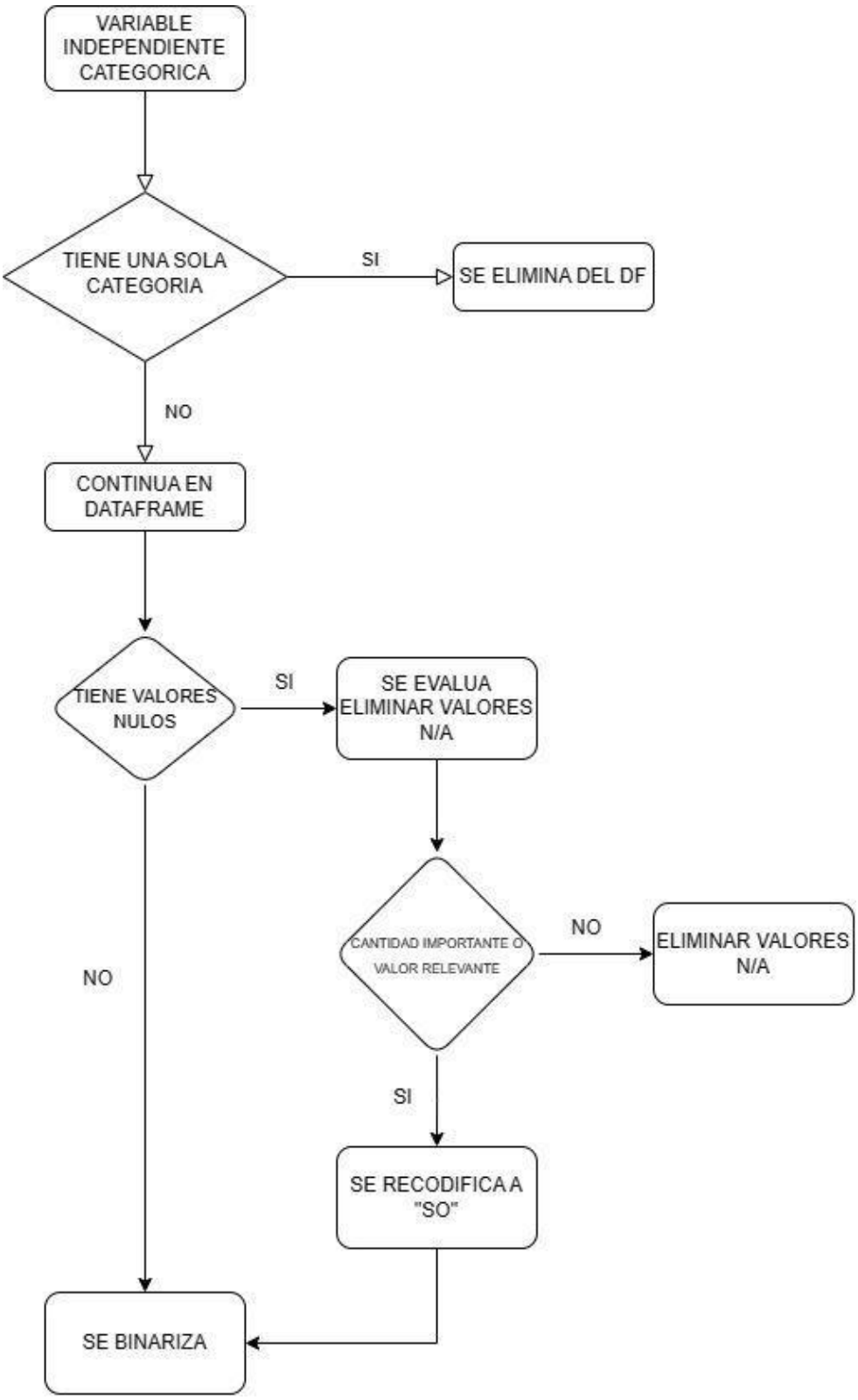

Figura 14 Diagrama de Valores Nulos

Capítulo 5. Análisis exploratorio post Procesamiento

Posterior al procesamiento de los datos, se realiza un nuevo análisis exploratorio para ver el impacto de estos cambios en la data y también para poder realizar análisis multivariados ahora que la data ya se encuentra apta para trabajar con ella.

5.1 Análisis valores perdidos

Como se puede observar en la figura 15, posterior al procesamiento de los datos, el *df* cuenta con 1.355 filas con datos, 29 variables (que se encuentran definidas en el anexo A.3) de las cuales 15 contienen datos de tipo *int64* y 13 del tipo object y 1 category (rango etario). Ninguna presenta valores atípicos ni nulos.

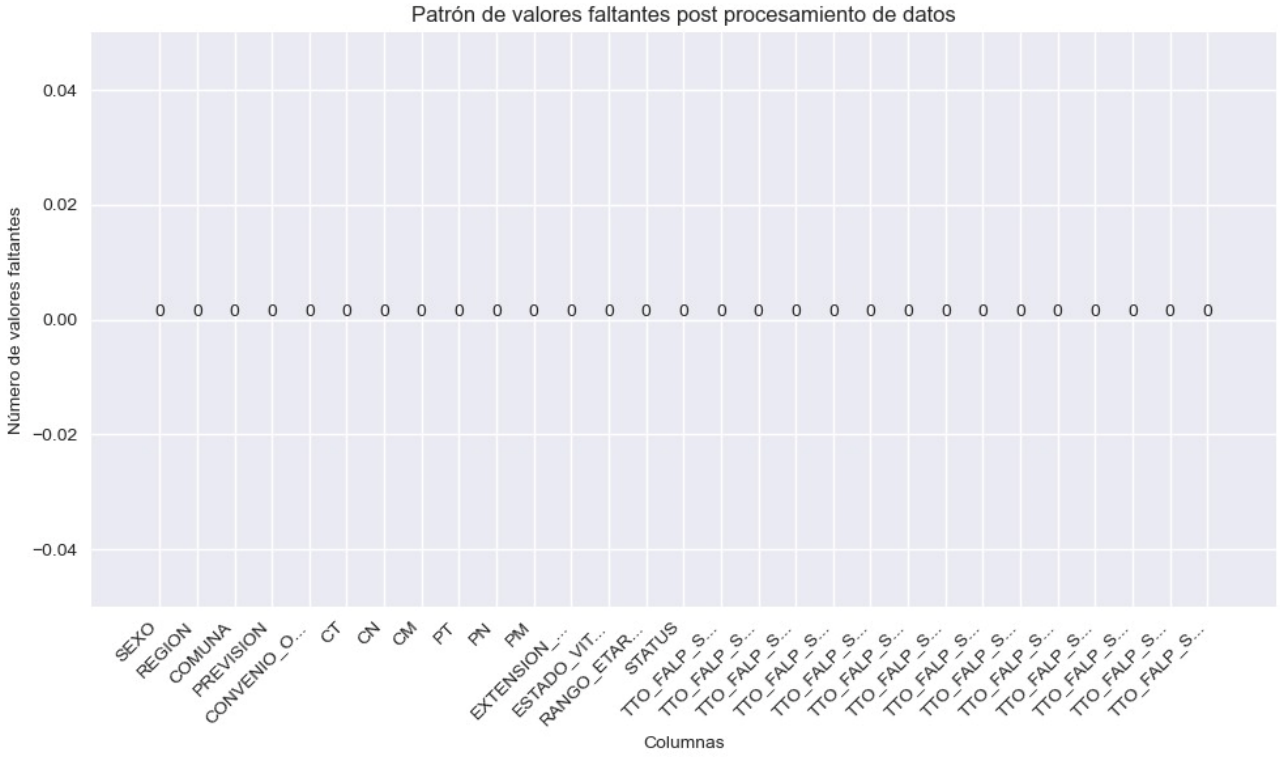

Figura 15 Diagrama de valores faltantes

Respecto a los cambios generales, en el caso de la variable “SEXO” la cantidad de hombres se redujo de 33 a 5 pacientes.

En el caso de la “Extensión del diagnóstico” sigue la misma distribución solamente variando la cantidad de pacientes en cada clasificación.

5.2 Análisis Univariado

A continuación, se presenta el análisis post procesamiento de todas las variables a analizar.

Estadios reclasificados a STATUS (I, II y III)

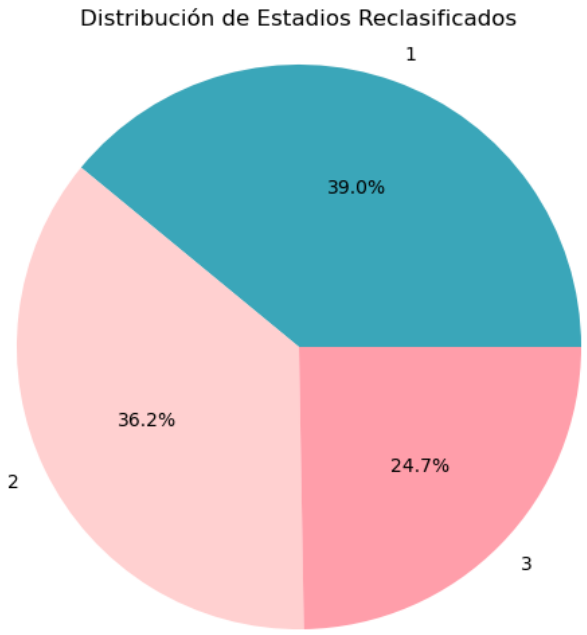

Figura 16 Distribución por tipo de Estadio reclasificado

La variable ESTADIO fue reclasificada, por lo que ahora se puede observar que la mayor parte de la muestra pertenece al STATUS 1 con un 39.04% indicando un cáncer en nivel “Inicial”, en el caso de STATUS 2 se ve representado con un 36.24% de pacientes con un cáncer en nivel “Intermedio” y por último los pacientes con un cáncer de mama en nivel “Avanzado” representan un 24.7% siendo estos los que deberían recibir atención y tratamiento de manera urgente.

Variables clínicas y patológicas post procesamiento

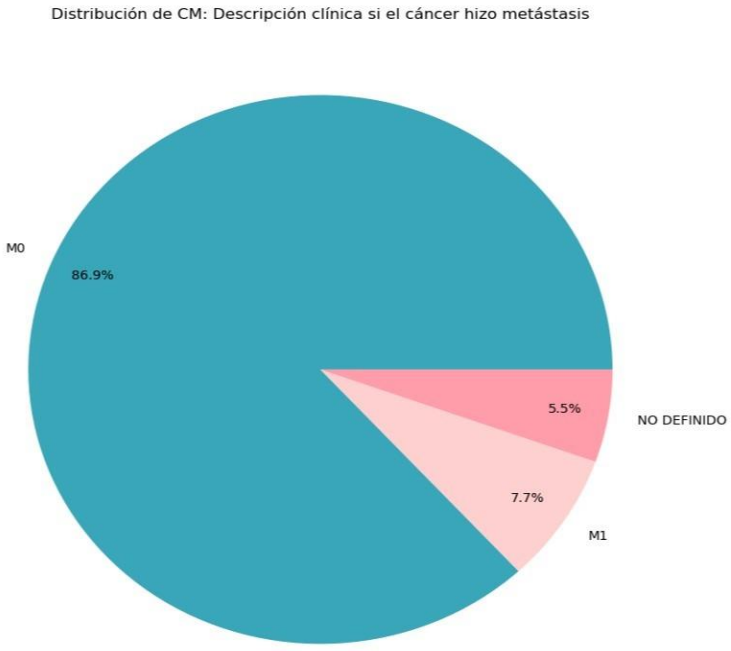

Figura 17 Descripción clínica metástasis del cáncer

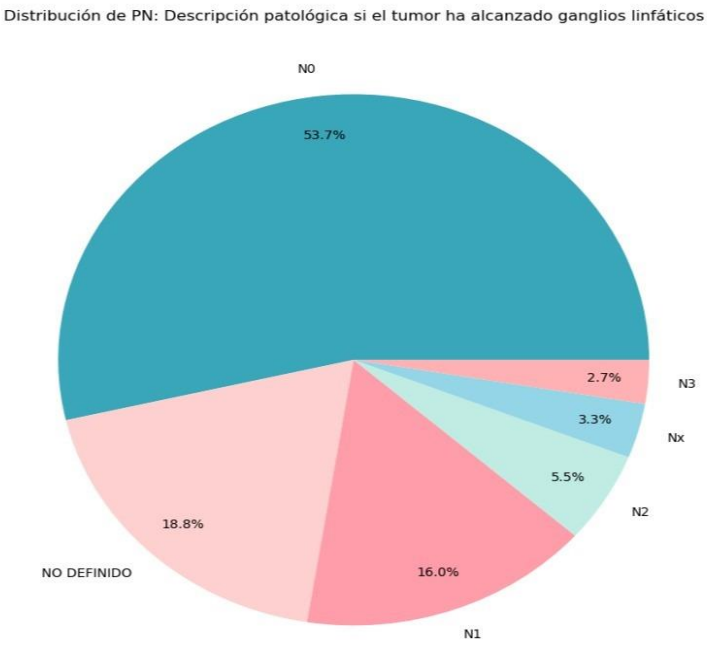

Figura 18 Descripción patológica ganglios linfático

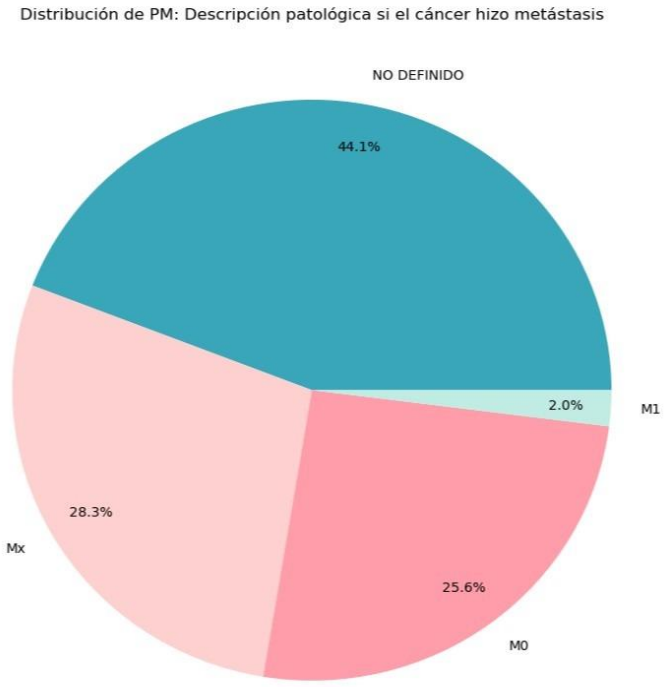

Figura 19 Descripción patológica metástasis cáncer

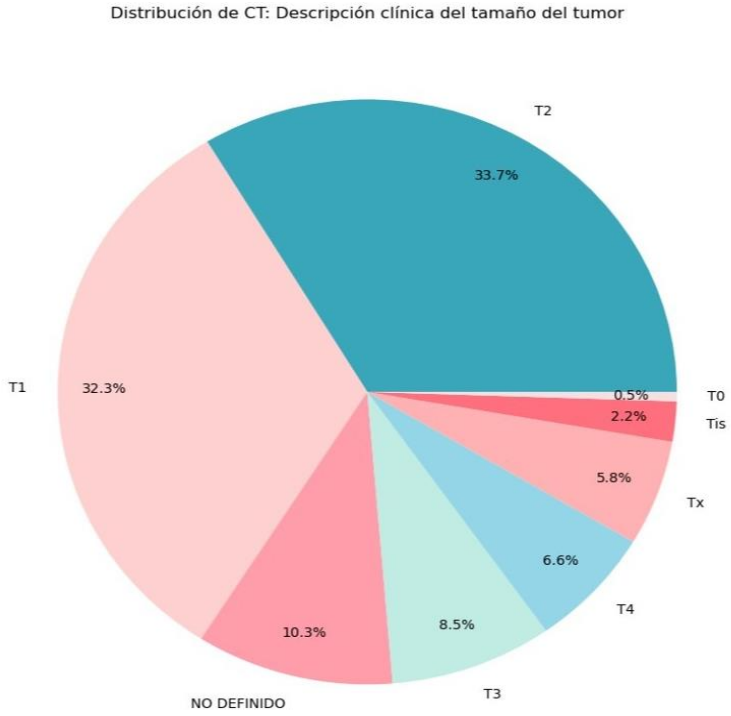

Figura 20 Descripción clínica del tamaño del tumor

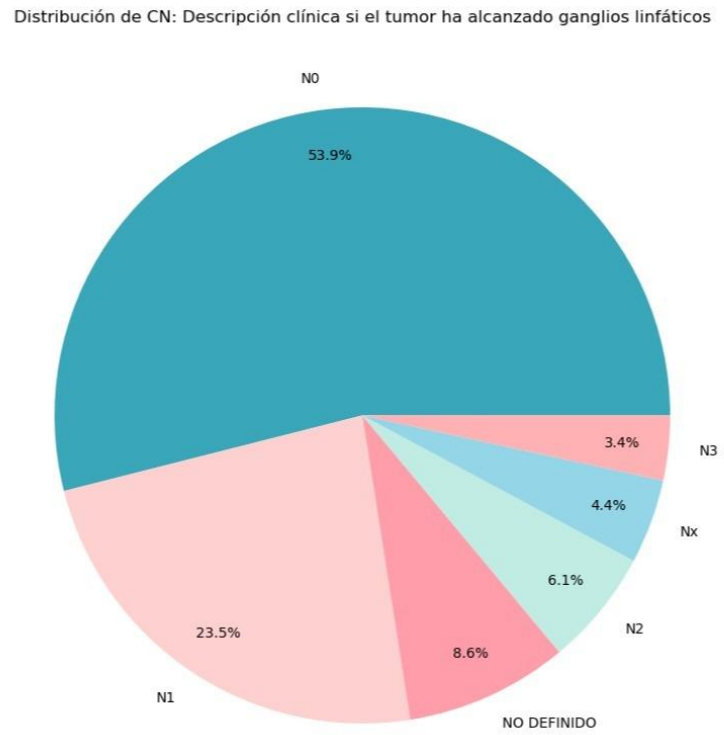

Figura 21 Descripción clínica ganglios linfáticos

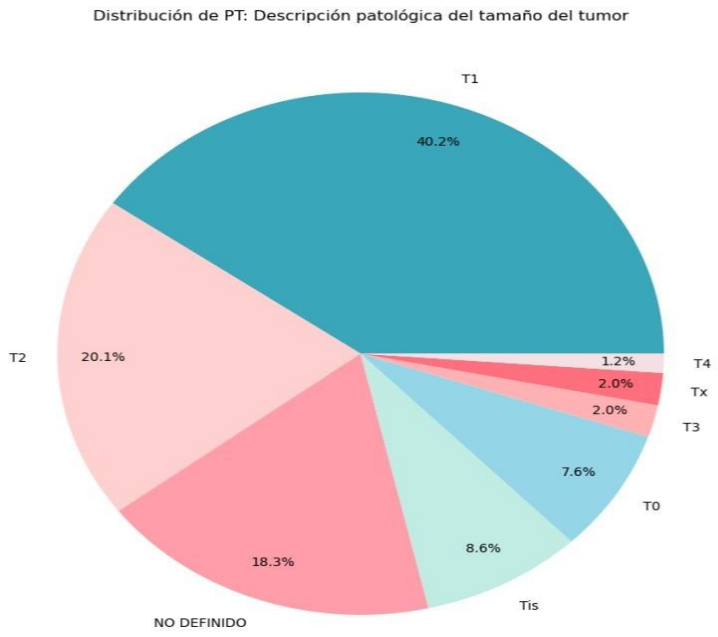

Figura 22 Descrpción patológica tamaño del tumor

Se puede apreciar que, en las seis variables, los datos nulos se reducen (ahora llamados “NO DEFINIDO”), resaltando así los parámetros de cada variable. Estos siguen la misma tendencia que tenían antes del procesamiento prevaleciendo las mismas conclusiones del análisis anterior.

Variable convenio oncológico

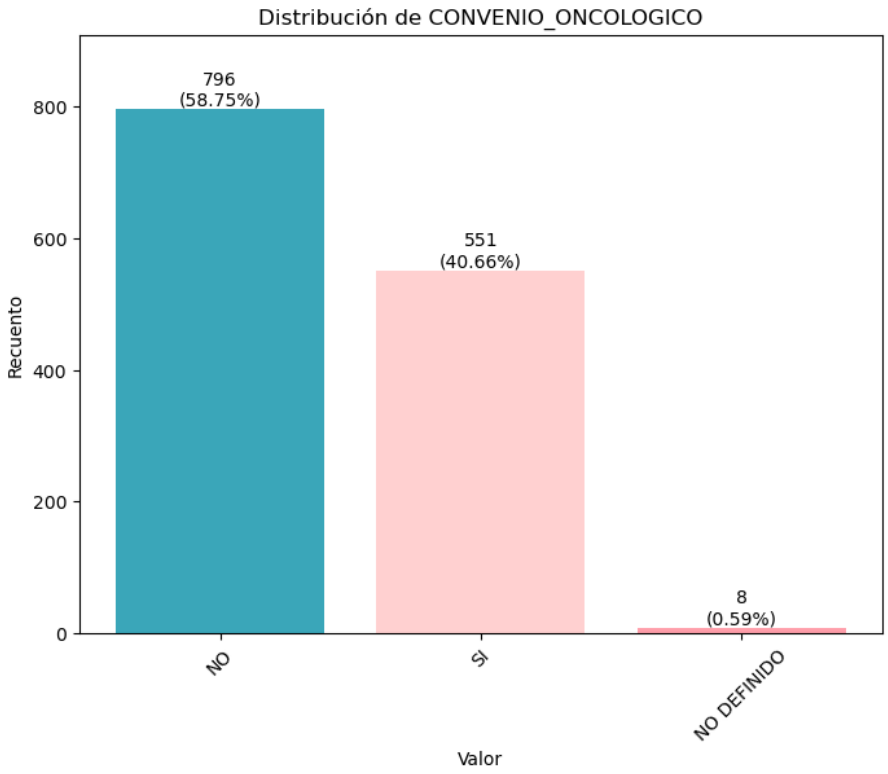

Figura 23 Distribución de convenio oncológico en los pacientes

En el gráfico de la figura 19, podemos observar que posterior al procesamiento de los datos, la moda de la variable “Convenio Oncológico” se invierte indicando que un 58.75% de los pacientes no son beneficiarios de la FALP.

Variable previsión

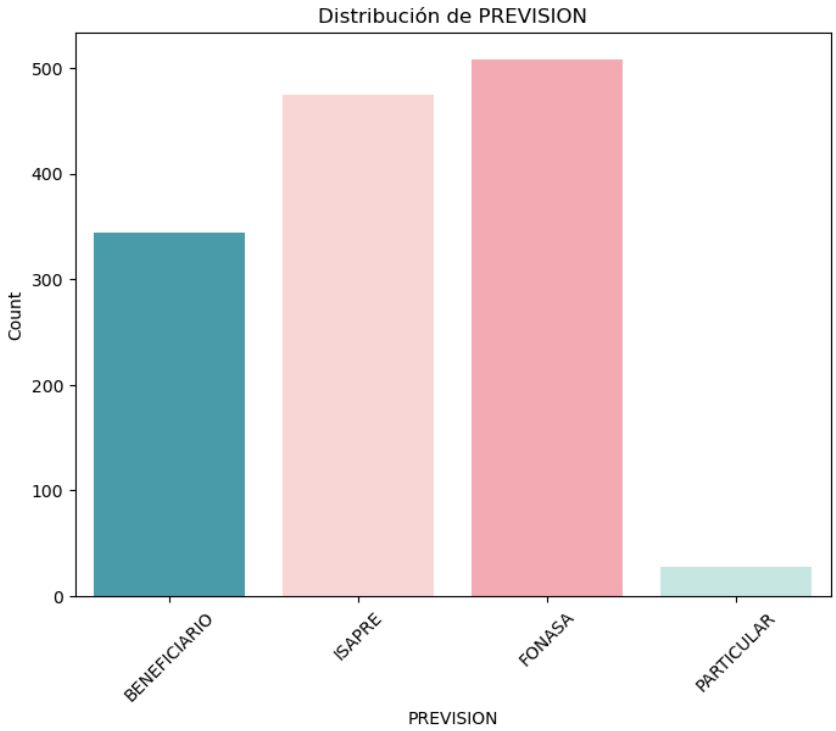

Figura 24 Distribución de la previsión que pertenecen los pacientes

Se observa que ahora la diferencia entre la cantidad de pacientes con distintas previsiones es menor, y que la mayor parte de los pacientes pertenecen al sistema de salud público de Chile FONASA.

Variable rango etario

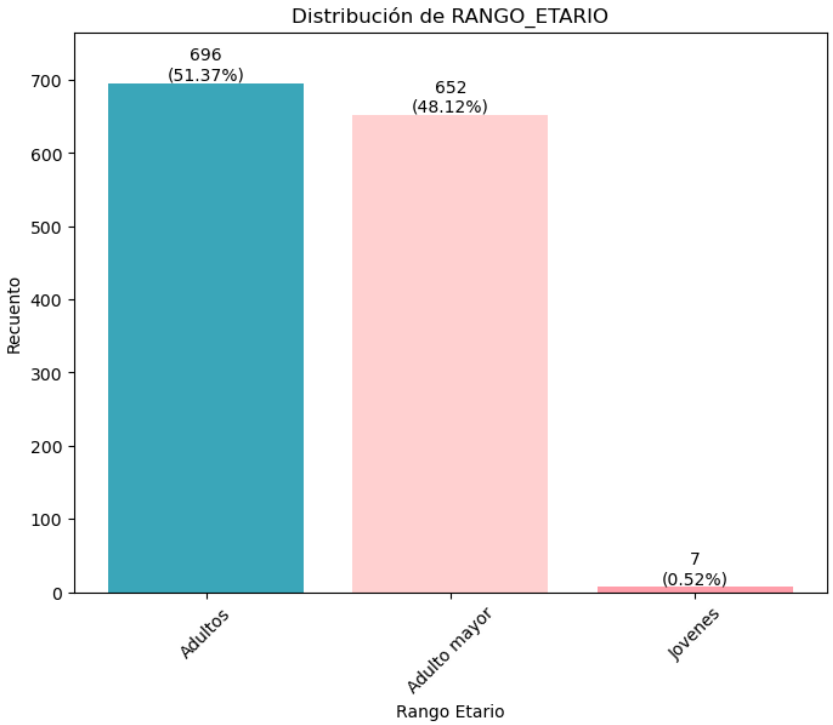

Figura 25 Distribución de rango etario

Respecto a la variable “Rango Etario” se puede observar que la mayor parte de la muestra se divide entre “Adultos” con un 51.37% correspondiente a personas entre 30 y 59 años, mientras que la otra mitad está compuesta por un 48.12% de “Adulto Mayor” correspondientes a personas de 60 años o más. Sólo hay 7 jóvenes presentes en la data (menos de 30 años).

Análisis multivariado

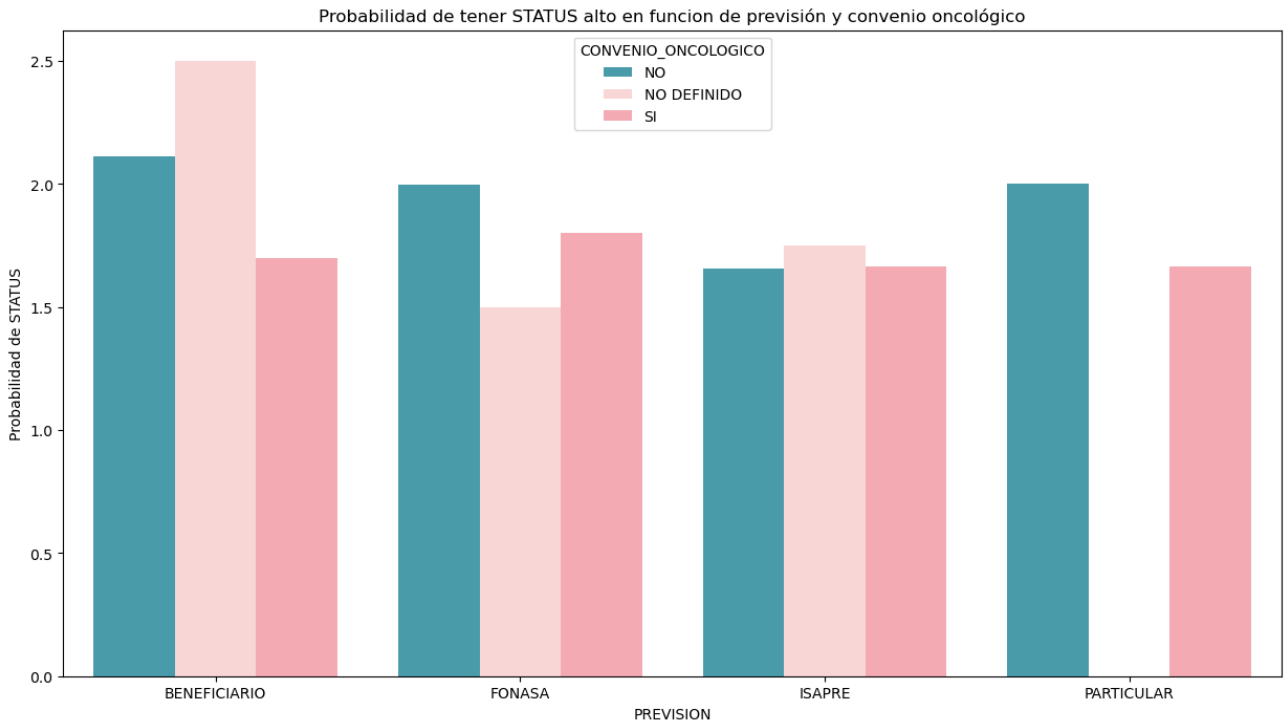

Figura 26 Previsión y Convenio Oncológico vs Status

Al analizar estas dos variables vs la probabilidad de tener Status Alto, se puede observar que los pacientes que tienen como previsión ISAPRE tienen una menor probabilidad de obtener un STATUS alto, es decir, de tener un cáncer Avanzado independientemente si cuentan con un convenio oncológico o no. Respecto a los pacientes que tienen como previsión Fonasa, Particular o Beneficiario, se observa que los que no cuentan con un Convenio Oncológico tienden a tener una mayor probabilidad de tener un cáncer Avanzado.

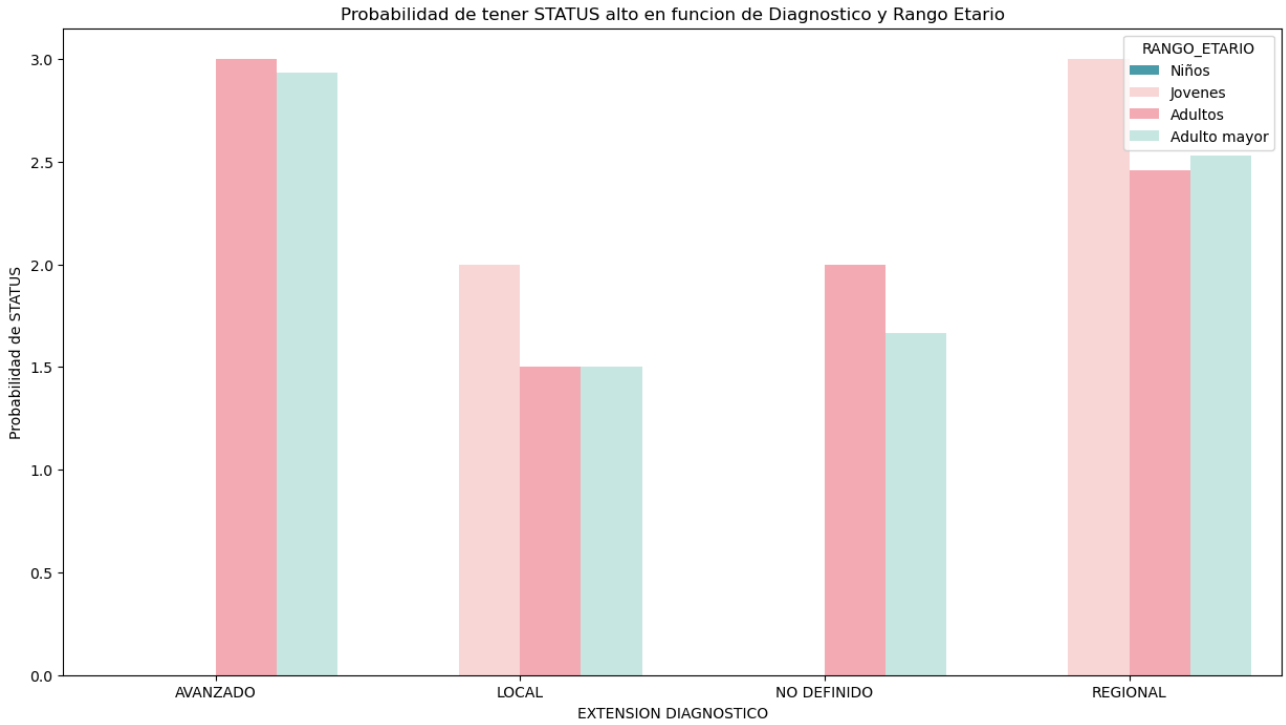

Figura 27 Rango Etario y Extensión del Diagnóstico vs Status

Al analizar estas dos variables vs la probabilidad de tener Status Alto destaca que los jóvenes tienen una alta probabilidad de tener STATUS alto con un diagnóstico Local o Regional mientras que nulas de tener un cáncer Avanzado. Esto puede deberse también a la poca cantidad de jóvenes con esta enfermedad.

Respecto a los Adultos, se puede observar que presentan una mayor tendencia a tener un Status Alto cuando el cáncer ya se ha ramificado en otras partes del cuerpo, esta situación se repite en el caso de los Adultos Mayores.

También destacar que cuando el cáncer se encuentra solamente en la mama (Local) la probabilidad de tener Status alto es baja.

Capítulo 6. Modelación y predicción del proyecto

6.1. Introducción

A continuación, se presenta un problema de clasificación multiclase supervisado. Esto, ya que se quiere predecir la urgencia de atención (**STATUS**) de los pacientes según variables clínicas, patológicas y sociodemográficas.

Dado lo anterior, se entrenan los siguientes modelos de clasificación multiclase y se hará una búsqueda de hiperparámetros que permitan encontrar los modelos más óptimos. Se utilizarán los modelos recién mencionados, ya que permiten la predicción de multiclases. Se utilizarán modelos menos costosos (**Bayes Ingenuo Multinomial y Árbol de Clasificación Simple)** a más costoso computacionalmente (**Random Forest de Clasificación, Gradient Boosting de Clasificación** y **Support Vector Machine**).

Para entrenar los modelos se utilizará el 66.7% de la data preprocesada. El otro 33.3% de los datos preprocesados serán utilizados para evaluar el desempeño de cada modelo. Las métricas que se utilizarán para medir el desempeño de cada modelo serán la exactitud, recall y accuracy para problemas de clasificación.

6.2. Bayes Ingenuo Multinomial

- **alpha:** Permite modificar qué tan suave es la curva de probabilidades del modelo. Se utilizarán los siguientes valores: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000].
- **fit_prior:** Permite aprender sobre la frecuencia de cada clase antes de realizar el modelo. Se utilizarán los siguientes valores: [True, False].

6.3. Árbol de clasificación simple

- **max_depth:** Se modificará este parámetro, ya que modifica cuantos niveles puede tener un árbol. Un árbol con muchos niveles puede incurrir en el overfitting. Se utilizarán los siguientes valores: [3, 4, 5, 6, 7, 8, 9, 10].
- **max_leaf_nodes:** Se modificará este parámetro, puesto que permite controlar la cantidad de nodos hoja. Esto evitará el overfitting. Se utilizarán los siguientes valores: [10, 50, 100].
- **criterion**: Este parámetro permite cambiar el modo en el que se calcula la impureza de una división. Se utilizarán los siguientes valores: [‘gini’, ‘entropy’].
- **min_samples_split**: Representa el número mínimo de muestras requeridas para que un nodo interno se divida en hijos. Se utilizarán los siguientes valores: [0.05, 2]

6.4. Regresión logística

- **penalty**: La regularización es una técnica utilizada para evitar el sobreajuste en modelos de aprendizaje automático. El parámetro "penalty" determina el tipo de regularización aplicada en la regresión logística
- **C**: Es un parámetro de regularización que controla el equilibrio entre lograr la máxima separación de los puntos de datos y permitir errores en la clasificación. Un valor más bajo de C permite una mayor cantidad de errores, lo que puede conducir a una frontera de decisión más suave, mientras que un valor más alto busca una mayor precisión en la clasificación. Se utilizarán los siguientes valores: [0.01, 0.1, 1.0, 10.0]
- **fit_intercept**: Es un parámetro booleano que determina si se debe ajustar o no el término de intercepción en el modelo de regresión logística. Si se establece en "True", se ajusta un término de intercepción, mientras que si se establece en "False", no se ajusta ningún término de intercepción. Se utilizarán los siguientes valores: [**True**, **False**],
- **Solver**: El parámetro "solver" especifica el algoritmo utilizado para optimizar los coeficientes en la regresión logística. Se utilizarán los siguientes valores: ["liblinear", "saga"]
- **max_iter**: Es el número máximo de iteraciones permitidas para que el algoritmo converja. Define cuántas veces se repite el proceso de optimización antes de que se detenga. Se utilizarán los siguientes valores: [100, 200, 300].

Los modelos 6.2, 6.3 y 6.4 Relacionados con:

- Regularización (ej: C, penalty)
- Modular Overfitting (ej: max_depth, max_leaf_nodes)

6.5. Random Forest de clasificación

- **max_depth**: Es el parámetro que controla la profundidad máxima de los árboles en el bosque. Limitar la profundidad puede evitar el sobreajuste al conjunto de entrenamiento. Un valor más alto indica árboles más profundos y complejos. Se utilizarán los siguientes valores: [5, 6, 7, 8, 9, 10].
- **max_leaf_nodes:** Se modificará este parámetro, puesto que permite controlar la cantidad de nodos hoja. Esto evitará el overfitting. Se utilizarán los siguiente valores: [10, 50, 100].
- **n_estimators**: Indica el número de árboles que se van a utilizar en el bosque. Cuanto mayor sea el número de estimadores, más precisa puede ser la clasificación, pero también aumentará el tiempo de entrenamiento y predicción. Se utilizarán los siguientes valores: [500, 600, 700, 800, 900, 1000].

6.6. Gradient Boosting de clasificación

- **max_depth**: Al igual que en el caso de Random Forest, max_depth controla la profundidad máxima de los árboles de decisión en el conjunto de clasificadores débiles. Un valor más alto permite modelos más complejos, pero puede aumentar el riesgo de sobreajuste. Se utilizarán los siguientes valores: [5, 6, 7, 8, 9, 10].
- **max_leaf_nodes:** Se modificará este parámetro, puesto que permite controlar la cantidad de nodos hoja. Esto evitará el overfitting. Se utilizarán los siguientes valores: [10, 50, 100].
- **Learning_rate**: Es el parámetro que controla la contribución de cada clasificador débil al modelo final. Un valor más bajo significa que cada clasificador tiene una contribución más pequeña, lo que hace que el proceso de entrenamiento sea más lento pero generalmente mejora la precisión del modelo. Se utilizarán los siguientes valores: [0.001, 0.01, 0.1, 1].

6.7. Support Vector Machine

- **C**: Es un parámetro de regularización que controla el equilibrio entre lograr la máxima separación de los puntos de datos y permitir errores en la clasificación. Un valor más bajo de C permite una mayor cantidad de errores, lo que puede conducir a una frontera de decisión más suave, mientras que un valor más alto busca una mayor precisión en la clasificación. Se utilizarán los siguientes valores: [0.1, 1, 10, 100].
- **kernel**: Especifica el tipo de función kernel a utilizar en el SVM. Los tipos comunes de kernel incluyen lineal, polinomial, radial (RBF) y sigmoide, entre otros. El kernel determina cómo se mapean los puntos de datos en un espacio de mayor dimensionalidad para permitir una separación no lineal de las clases. Se utilizarán los siguientes valores: [‘linear’, ‘poly’].
- **degree**: Este parámetro sólo es relevante cuando se utiliza un kernel polinomial. Indica el grado del polinomio a utilizar en la función de kernel. Un grado más alto permite un modelo más complejo y flexible, pero también puede aumentar el riesgo de sobreajuste. Se utilizarán los siguientes valores: [2, 3, 4].

Los modelos 6.5, 6.6 y 6.7 relacionados con:

- Regularización (ej: C)
- Modular Overfitting (ej: max_depth, max_leaf_nodes, n_estimators)

6.8. Resultados

Tabla 3 Métricas de desempeño

Modelo	Accuracy Grid Search	Recall STATUS 1	Recall STATUS 2	Recall STATUS 3	Validación Accuracy
Regresión Logística	0.91	0.94	0.95	0.90	0.93
Naive Bayes	0.85	0.92	0.86	0.83	0.88
Decision Tree Classifier	0.89	0.96	0.95	0.87	0.93
Random Forest	0.90	0.96	0.95	0.85	0.93
Gradient Boosting	0.93	0.96	0.90	0.95	0.94
Support Vector Machine	0.90	0.94	0.93	0.89	0.92

En la tabla N°3, se puede observar que la mayoría de los modelos (a excepción de Naive Bayes), presentan desempeños superiores a 0.9.

Las métricas de recall (1,2 y 3) muestran al modelo Gradient Boosting como el mejor de todos. Esto se deduce de las métricas que presenta en recoll en las clases 1 y 2. En cuanto a la clase 3, tanto en Gradient Boosting como en Support Vector Machine, se presentan las mismas métricas. En accuracy de validación, se observa que el modelo Gradient Boosting, presenta el mayor valor, alcanzando un 0.94.

En base a los resultados obtenidos por los modelos propuestos, se decidió implementar el modelo Gradient Boosting Classifier debido a sus métricas en el Recall. Este modelo presenta un 5% más de acierto en la variable STATUS 3 que corresponde a los pacientes de mayor gravedad.

6.9. best_estimator

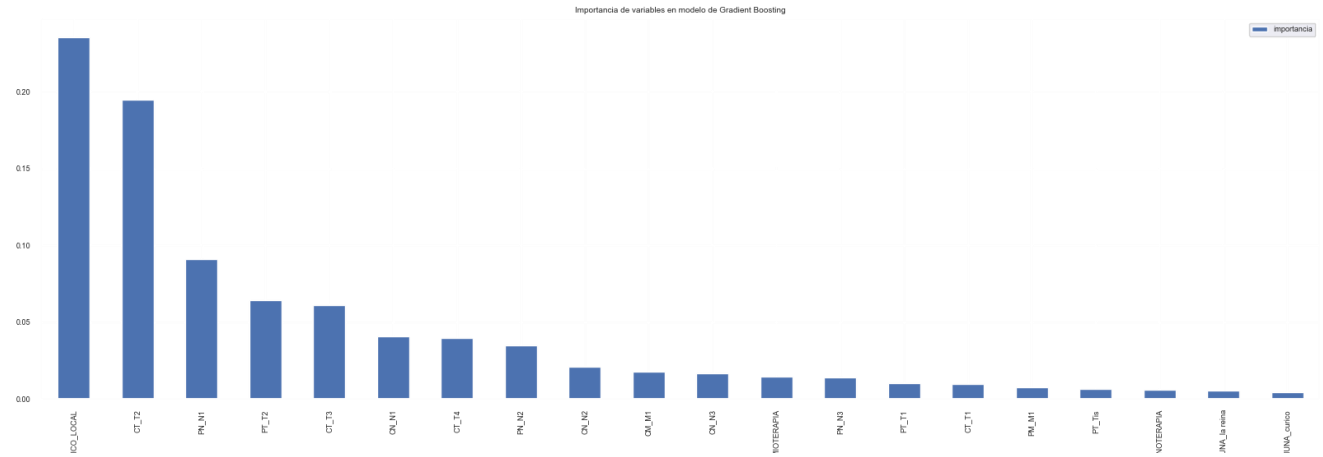

Figura 28 Gráfico de variables más influyentes en el modelo Gradient Boosting Classifier

En el presente gráfico, se puede observar que las variables más influyentes en el modelo de Regresión logística son: Extensión diagnóstica, CT_T2 y PN_N1.

La localización y tamaño del tumor están directamente relacionados con el estadio del cáncer. De acuerdo al gráfico, la localización del tumor (Local, regional y avanzado) es un indicador importante e influyente en establecer que tan inicial o avanzado está el estadio del cáncer.

6.10 Grafico Regresión Logística

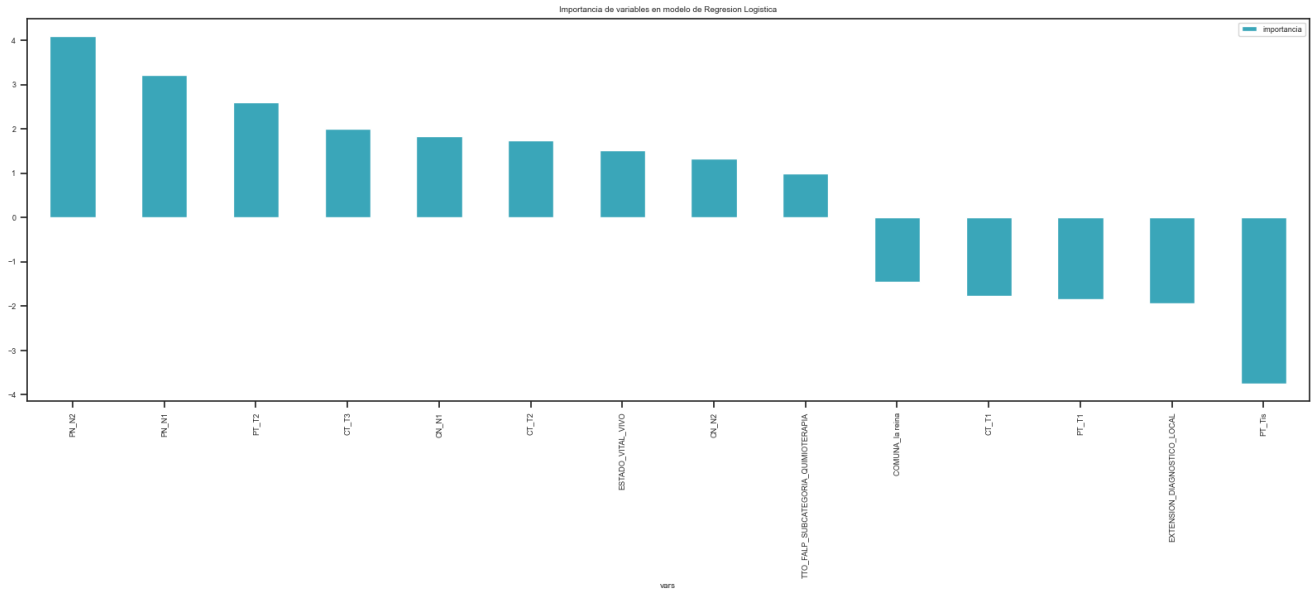

Figura 29 Grafico Regresión Logística

En la Figura 31, se puede observar que:

Las variables ubicadas sobre el eje Y (> 0) predicen un cáncer en estadio avanzado (II, III o IV), cuando:

- Existe presencia de cáncer en ganglios linfáticos (PN, CN > N1)
- Tamaño del tumor (PT, CT > T2)

Las variables ubicadas bajo el eje Y (<0) predicen un cáncer en estadio leve (0 o I):

- Sin diseminación (PT_Tis, Extensión local)
- Tamaño pequeño (nivel T1)

6.11 Árbol de clasificación

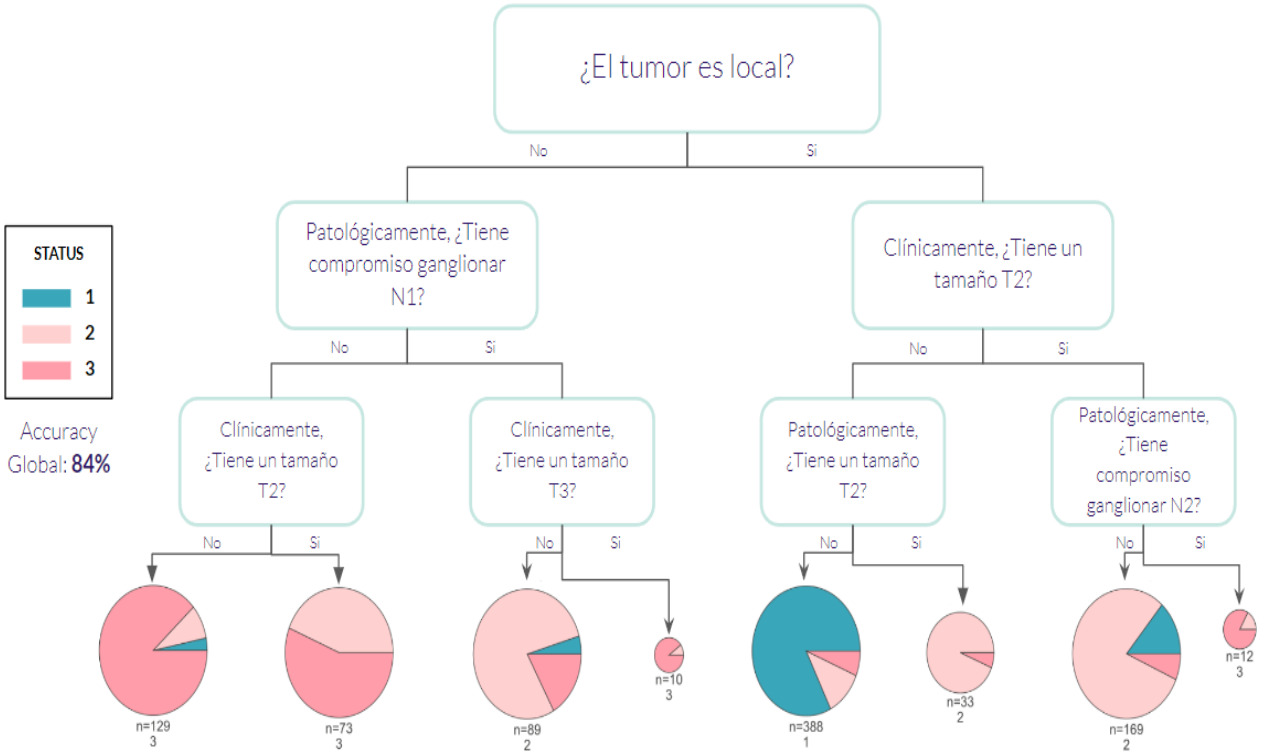

Figura 30 Árbol de clasificación

El árbol de clasificación se puede interpretar de la siguiente manera:

- La primera bifurcación se realiza considerando si el tumor es de naturaleza local o diseminada (nodo raíz). Si el tumor es de carácter local, entonces se sigue por la rama derecha, mientras que si es diseminado se sigue por la rama izquierda.

- La mayor cantidad de tumores graves se encuentran en la rama más izquierda del árbol. Estos pacientes presentan un cáncer diseminado, tienen la presencia de tejido canceroso distinto al nivel N1 y pueden presentar un tamaño de nivel T2 o un nivel distinto.

- La mayor cantidad de tumores leves se encuentran en la rama derecha del árbol. Estos pacientes presentan un cáncer local y tienen tamaños distintos a T2 (tanto clínica, como Patológicamente).

- Un tumor en STATUS 2 puede seguir diversas ramas del árbol. Lo anterior indica que las características de un tumor en STATUS 2 son heterogéneas. Los resultados muestran que este tipo de tumor puede ser tanto local como diseminado, con compromiso ganglionar y tamaños variables.

Capítulo 7. Proyección y solución

7.1. Formulario Web

Con la finalidad de darle utilidad a la investigación y predicción realizada con este proyecto, se utilizará el modelo en un formulario web demo que servirá para ingresar los datos de un paciente y los resultados de sus exámenes clínicos o patológicos. Existen requisitos para el ingreso de estos datos procurando un ingreso limpio de la data y permitiendo mejores resultados en el modelo predictivo.

El TextBox de EDAD tiene límites sólo un valor numérico entre 0 y 99 años y es de carácter obligatorio. Los demás valores como PREVISION, REGION, COMUNA, EXTENSION_DIAGNOSTICO, TTO_FALP_SUBCATEGORIA y las descripciones TNM patológicas y clínicas ingresan por DropDown lo que permite que el valor seleccionado sea limpio y sin errores de tipeo. Variables como SEXO, CONVENIO_ONCOLOGICO ingresaran sus valores con la selección de RadioButton y CheckBox correspondientemente. La demo además no permite que los campos de descripción clínicas y patológicas del tumor ingresen vacíos, al menos uno de ellos debe ingresar con algún valor.

https://status-cancer.anvil.app/

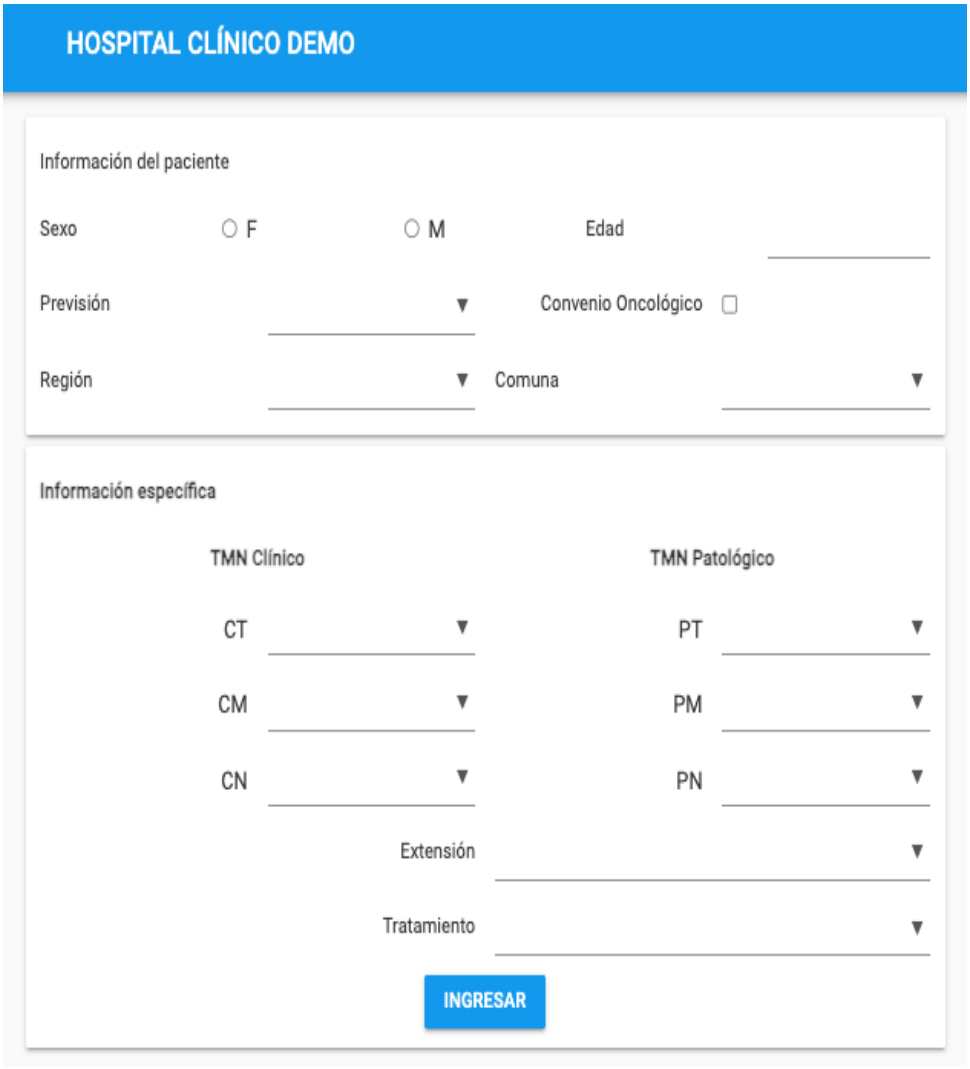

Figura 31 Formulario app

La app fue programada como mockup con la herramienta Anvil y un notebook en GoogleColab lo que permite que funcione como su propio servidor para montar la demo.

Al ingresar los datos y presionar el botón “Ingresar” los datos son completados y procesados para entrar en el modelo predictivo. Esta predicción es retornada a la app, la que evalúa si el valor de la predicción STATUS es igual a 3 o no, de cumplirse esta condición la app levanta una alarma donde se le informa al usuario que registra la información que el paciente recientemente ingresado cumple con las características de un ESTADIO III o IV por lo que debe ser atendido a la brevedad.

La finalidad de esta app es presentarse como una herramienta de priorización a la atención de los pacientes, colaborando con la declaración del cáncer que deben realizar los médicos oncólogos.

Capítulo 8. Conclusiones

A raíz de los resultados obtenidos en la implementación de los diferentes modelos de Machine Learning, se puede concluir que los modelos de aprendizaje automático son una herramienta poderosa para predecir el comportamiento que tendrá el cáncer de mama en el paciente.

Los datos clínicos que se administran durante toda la atención del paciente, permiten a través de un modelo predictivo de clasificación identificar patrones de comportamiento y realizar predicciones de estadio con una certeza superior al 95 %.

Una correcta predicción de estadio está directamente relacionada con un buen registro e ingreso de los datos clínicos y patológicos del paciente, por lo que es indispensable proponer el uso de una aplicación que permita ingresar datos estructurados y ordenados por cada atención realizada. Lo anterior, permite trabajar con la mayor cantidad de data disponible, disminuyendo el ingreso de datos nulos y/o vacíos.

La predicción del estadio permite alertar a los gestores de casos oncológicos a través de una alarma que les indique cuando un paciente es considerado de riesgo. Esta alarma permite a los clínicos, priorizar la atención, facilitando una detección temprana y más precisa del cáncer, lo que otorga al paciente la posibilidad de acceder a un tratamiento adecuado y fomentar una atención preventiva en salud.

Dado lo anterior, se puede concluir, que una herramienta de Machine Learning, junto a una aplicación web que permita estandarizar y poner a disposición los datos de atención de un paciente, pueden predecir el estadio del cáncer con una precisión del 95 %, siendo una herramienta de apoyo confiable para el clínico, permitiendo generar data de calidad y facilitando la clasificación del estadio. Esta herramienta puede ser un apoyo en los diferentes niveles de atención pública y privada del país y un paso hacia las mejoras en salud digital.

Capítulo 9. Referencias

[1] Global Cancer Observatory, [En línea]. Available: https://gco.iarc.fr/

[2] Departamento de Estadísticas e Información en Salud, Gobierno de Chile, [En línea]. Available: https://informesdeis.minsal.cl/

Anexos

A.1 Definición y clasificación de Estadio.

El estadio de un tumor por cáncer de mama se determina en función de sus características, como su tamaño, y si tiene o no receptores hormonales. Conocer el estadio del cáncer permite:

- Determinar el pronóstico, es decir, el resultado probable del tratamiento del cáncer de mama.
- Elegir las mejores opciones de tratamiento.
- Definir si algunos estudios clínicos pueden aportar nueva información al tratamiento.

Los estadios del cáncer con los siguientes:

- **Estadio 0:** Corresponde a los tipos de cáncer no invasivos que permanecen en su ubicación original.
- **Estadio I:** El estadio I describe el cáncer de mama invasivo (las células cancerosas toman o invaden el tejido mamario normal que las rodea). El estadio I está dividido en subcategorías, conocidas como “IA” y “IB”.
- **Estadio II:** Cáncer de mama invasivo. El estadio II se divide en las subcategorías IIA y IIB.
- **Estadio III:** Cáncer de mama invasivo. El estadio III se divide en las subcategorías IIIA, IIIB y IIIC.
- **Estadio IV:** Corresponde a los tipos de cáncer invasivos que hacen metástasis fuera de la mama en otras partes del cuerpo.

En informes patológicos: Se determina si se limita a una zona de la mama o si hizo metástasis en tejidos sanos dentro de la mama u otras partes del cuerpo.

Durante la cirugía de extracción se observa uno o más de los ganglios linfáticos ubicados en la axila, que es donde tiende a trasladarse primero el cáncer de mama.

En análisis de sangre o en estudios por imágenes (Ecografías, Mamografías, RM de mamas).

El sistema de estadificación del cáncer de mama, conocido como "sistema TNM", está regulado por el Comité Conjunto Estadounidense sobre el Cáncer (AJCC, sigla en inglés)

- **T**: El tamaño del tumor y si hizo metástasis en tejido cercano o no.
- **N**: Si hay un tumor en los ganglios linfáticos.
- **M**: Si el cáncer hizo metástasis en otras partes del cuerpo.

Mientras más altos los números, más avanzado está el cáncer.

La categoría T (tamaño) describe el tumor original (primario):

- **TX**: significa que el tumor no puede evaluarse.
- **T0**: significa que no hay indicio alguno del tumor primario.
- **Tis**: Significa que el tumor se encuentra in situ (es decir, no comenzó a avanzar hacia tejido mamario sano).
- **T1, T2, T3, T4**: Estos números se basan en el tamaño del tumor y en qué medida ha crecido dentro del tejido mamario circundante. Cuanto más alto es el número T, más grande es el tumor o más pudo haber crecido en el tejido mamario.

La categoría N (afectación de los ganglios linfáticos) indica si el cáncer ha llegado a los ganglios linfáticos cercanos:

- **NX**: Significa que los ganglios linfáticos próximos no pueden evaluarse, por ejemplo, si se extirparon con anterioridad.
- **N0**: significa que los ganglios linfáticos cercanos no presentan cáncer.
- **N1, N2, N3**: Estos números se basan en la cantidad de ganglios linfáticos afectados y qué nivel de cáncer presentan. Cuanto más alto es el número N, mayor es el nivel de afectación de los ganglios linfáticos.

La "M" del sistema TNM describe si el cáncer se ha diseminado a otras partes del cuerpo, lo que se denomina "metástasis". Esto supone que ya no se considera un cáncer de avance local o en estadio temprano.

- **MX:** No se puede evaluar la diseminación a distancia.
- **M0**: No hay evidencia clínica de metástasis distantes.

- **M0 (i+):** no hay evidencia clínica ni radiográfica de metástasis distantes. Sin embargo, la evidencia microscópica de células tumorales se encuentra en la sangre, la médula ósea u otros ganglios linfáticos que no midan más de 0.2 mm.
- **M1**: hay evidencia de metástasis en otra parte del cuerpo, es decir, hay células mamarias cancerosas que crecen en otros órganos.

Estadios del cáncer y su disipación:

- **Local**: El cáncer está restringido a la mama.
- **Regional**: Los ganglios linfáticos han sido afectados, especialmente aquellos ubicados en la axila.
- **Distante**: El cáncer también se encuentra en otras partes del cuerpo.

A.2 Variables del Data Set

Variable	Descripción
ID_CASO	ID del paciente.
CATEGORIA	Categoría diagnóstica del tumor (ej.: órgano digestivo).
SUBCATEGORIA	Subcategoría diagnóstica del tumor (ej.: colon, páncreas o estómago).
SEXO	Sexo del paciente. Opciones: F o M, Femenino o Masculino respectivamente.
EDAD	Edad del paciente.
REGIÓN	Región del país de residencia.
COMUNA	Comuna o condado de residencia.
PREVISIÓN	Seguro médico del paciente. Opciones: FONASA (seguro médico público), BENEFICIARIO (beneficiario, subconjunto de FONASA), ISAPRE (seguro médico privado), PARTICULAR (sin seguro médico), FFAA (seguro médico militar).
CONVENIO_ONCOLOGICO	Beneficiario FALP. Opciones: SI, NO.
FECHA_DIAGNOSTICO	Fecha de diagnóstico.
CT	Descripción clínica del tamaño del tumor.
CN	Descripción clínica si el tumor ha alcanzado ganglios linfáticos.
CM	Descripción clínica si el cáncer hizo metástasis.
PT	Descripción patológica del tamaño del tumor.
PN	Descripción patológica si el tumor ha alcanzado ganglios linfáticos.
PM	Descripción patológica si el cáncer hizo metástasis.
ESTADIO	Estadio del cáncer en la fecha de diagnóstico. Opciones: O, I, II, III, IV.
EXTENSION_DIAGNOSTICA	Extensión del tumor en la fecha de diagnóstico. Opciones: LOCAL, REGIONAL, AVANZADO, PERITONEAL.
ESTADO_VITAL	Estado vital del paciente. Opciones: VIVO, FALLECIDO.
FECHA_DEFUNCION	Fecha de fallecimiento.
TTO_FALP_SUBCATEGORIA	Subcategoría del tratamiento.
FECHA_INICIO_TTO	Fecha de inicio del tratamiento.

FECHA_FIN_TTO	Fecha de fin del tratamiento.

A.3 Variables del Data Set Procesado

Variable	Descripción
SEXO	Sexo del paciente. Opciones: F o M, Femenino o Masculino respectivamente.
REGION	Región del país de residencia.
COMUNA	Comuna o condado de residencia.
PREVISION	Seguro médico del paciente. Opciones: FONASA (seguro médico público), BENEFICIARIO (beneficiario, subconjunto de FONASA), ISAPRE (seguro médico privado), PARTICULAR (sin seguro médico), FFAA (seguro médico militar).
CONVENIO_ONCOLOGICO	Beneficiario FALP. Opciones: SI, NO.
CT	Descripción clínica del tamaño del tumor.
CN	Descripción clínica si el tumor ha alcanzado ganglios linfáticos.
CM	Descripción clínica si el cáncer hizo metástasis.
PT	Descripción patológica del tamaño del tumor.
PN	Descripción patológica si el tumor ha alcanzado ganglios linfáticos.
PM	Descripción patológica si el cáncer hizo metástasis.
EXTENSION_DIAGNOSTICO	Extensión del tumor en la fecha de diagnóstico. Opciones: LOCAL, REGIONAL, AVANZADO, PERITONEAL.
ESTADO_VITAL	Estado vital del paciente. Opciones: VIVO, FALLECIDO.
RANGO_ETARIO	Rango etario del paciente.
STATUS	Agrupación según ESTADIO del cáncer del paciente.
TTO_FALP_SUBCATEGORIA_BP EXCISIONAL, BP INCISIONAL O AMPLIACIÓN DE MÁRGENES'	Indicador si el paciente se ha sometido a tratamiento BP EXCISIONAL, BP INCISIONAL O AMPLIACIÓN DE MÁRGENES.
TTO_FALP_SUBCATEGORIA_CIRUGÍA	Indicador si el paciente se ha sometido a tratamiento CIRUGÍA.
TTO_FALP_SUBCATEGORIA_CUIDADOS PALIATIVOS	Indicador si el paciente se ha sometido a tratamiento CUIDADOS PALIATIVOS.
TTO_FALP_SUBCATEGORIA_HORMONOTERAPIA	Indicador si el paciente se ha sometido a tratamiento HORMONOTERAPIA.

TTO_FALP_SUBCATEGORIA_INMUNOTERAPIA	Indicador si el paciente se ha sometido a tratamiento INMUNOTERAPIA.
TTO_FALP_SUBCATEGORIA_INMUNOTERAPIA + TERAPIAS MOLECULARES	Indicador si el paciente se ha sometido a tratamientos INMUNOTERAPIA y TERAPIAS MOLECULARES.
TTO_FALP_SUBCATEGORIA_NO DEFINIDO	Indicador si el paciente no tiene definido algún tratamiento.
TTO_FALP_SUBCATEGORIA_QUIMIOTERAPIA	Indicador si el paciente se ha sometido a tratamiento QUIMIOTERAPIA.
TTO_FALP_SUBCATEGORIA_QUIMIOTERAPIA + INMUNOTERAPIA	Indicador si el paciente se ha sometido a tratamientos QUIMIOTERAPIA E INMUNOTERAPIA.
TTO_FALP_SUBCATEGORIA_QUIMIOTERAPIA + INMUNOTERAPIA + TERAPIAS MOLECULARES	Indicador si el paciente se ha sometido a tratamientos QUIMIOTERAPIA, INMUNOTERAPIA y TERAPIAS MOLECULARES.
TTO_FALP_SUBCATEGORIA_QUIMIOTERAPIA + TERAPIAS MOLECULARES	Indicador si el paciente se ha sometido a tratamientos QUIMIOTERAPIA y TERAPIAS MOLECULARES.
TTO_FALP_SUBCATEGORIA_RADIOTERAPIA	Indicador si el paciente se ha sometido a tratamiento RADIOTERAPIA.
TTO_FALP_SUBCATEGORIA_RESECCIÓN ENDOSCÓPICA	Indicador si el paciente se ha sometido a tratamiento RESECCIÓN ENDOSCÓPICA.
TTO_FALP_SUBCATEGORIA_TERAPIAS MOLECULARES	Indicador si el paciente se ha sometido a tratamiento TERAPIAS MOLECULARES.