**DESAFÍO LATAM**

ACADEMIA DE TALENTOS DIGITALES



# Predicción de estadio del cáncer de mama

# usando herramientas de Machine Learning

**RESUMEN EJECUTIVO**

**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

# Tabla de Contenidos

[Capítulo 1. Introducción 1](#_30j0zll)

[1.1. Objetivos 2](#_2et92p0)

[Capítulo 2. Planificación de la investigación 2](#_2s8eyo1)

[2.1. Definición de la pregunta de investigación 3](#_17dp8vu)

[2.2. Hipótesis de trabajo 3](#_3rdcrjn)

[2.3. Definición del Vector Objetivo 3](#_26in1rg)

[2.4. Estrategias analíticas a nivel descriptivo 3](#_lnxbz9)

[Capítulo 3. Análisis exploratorio de datos iniciales 4](#_44sinio)

[Capítulo 4. Preprocesamiento de los datos 6](#_2jxsxqh)

[Capítulo 5. Análisis exploratorio post procesamiento 7](#_hetrkvuw3kgp)

[Capítulo 6. Modelación y predicción del proyecto 8](#_z337ya)

[Capítulo 7. Proyección y solución 9](#_2a5vkgwmvyje)

[Capítulo 8. Conclusiones 10](#_49x2ik5)

# Introducción

El cáncer se ha convertido en una de las principales causas de muerte a nivel mundial en las últimas dos décadas. En Chile, las neoplasias representan la segunda causa de muerte por enfermedades, y el cáncer de mama es la principal causa de muerte entre las mujeres en edad reproductiva. Con el objetivo de mejorar la atención y la prevención, se han implementado políticas públicas para facilitar el acceso a exámenes y especialistas.

Sin embargo, la falta de tabulación y estructuración de los datos de salud dificulta su análisis y la toma de decisiones informadas. Por lo tanto, este proyecto se enfoca en desarrollar un modelo que permita predecir el estadio del cáncer de mama utilizando datos patológicos y clínicos. Esta herramienta podría ayudar a los profesionales de la salud a priorizar la atención de los pacientes según la gravedad de su enfermedad.

Al mejorar la calidad de los datos y utilizar modelos predictivos, se espera generar una herramienta que pueda evaluar la agresividad del cáncer de mama. Esto permitiría una atención más precisa y oportuna, optimizando los recursos disponibles y mejorando los resultados para los pacientes.

Este proyecto busca utilizar datos estructurados para predecir el estadio del cáncer de mama y proporcionar a los referentes clínicos una herramienta adicional para priorizar la atención de los pacientes. Esto puede contribuir a un enfoque preventivo más efectivo y a mejores resultados en el tratamiento del cáncer de mama en Chile.

## Objetivos

**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, a partir de variables clínicas, con el fin de proporcionar una herramienta de priorización de pacientes y contribuir al ámbito clínico y de gestión.

**Objetivos Específicos**

* Realizar un análisis exploratorio de los datos para identificar patrones y relaciones entre las variables clínicas y los estadios del cáncer de mama.
* Diseñar y desarrollar un modelo predictivo utilizando algoritmos de Machine Learning para predecir los estadios del cáncer de mama.
* Evaluar y comparar el rendimiento de diferentes modelos de Machine Learning utilizando métricas adecuadas.
* Implementar una interfaz o herramienta accesible para ingresar los datos clínicos y obtener predicciones del estadio del cáncer de mama.
* Analizar el impacto potencial del modelo en la priorización de pacientes y su aplicación en la práctica clínica, considerando la optimización de recursos y la mejora en la detección temprana del cáncer de mama.

Este trabajo tiene como objetivo desarrollar un modelo predictivo para la detección temprana del cáncer de mama, con el fin de mejorar la priorización de pacientes y contribuir a la gestión clínica en esta área.

# Planificación de la investigación

## 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?

## 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.

## 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.**

## Estrategias analíticas a nivel descriptivo

Las estrategias analíticas a nivel descriptivo se centran en explorar y comprender los datos relacionados con el cáncer de mama. Esto implica realizar un análisis exploratorio de datos, caracterizar las variables utilizadas, analizar la distribución de los datos, investigar las correlaciones entre variables y, si es posible, segmentar los datos según características relevantes. Estas estrategias proporcionan una visión general de las variables, identifican patrones y tendencias, y sientan las bases para el desarrollo de un modelo de clasificación preciso.

# Análisis exploratorio de datos iniciales

En esta sección del texto, se lleva a cabo un análisis exploratorio detallado de los datos proporcionados por FALP sobre el cáncer de mama. Se enfoca en comprender y caracterizar la distribución de los datos, así como las relaciones existentes entre las variables relevantes.

Se destaca que el análisis se centra exclusivamente en el cáncer de mama debido a su importancia en términos de mortalidad a nivel nacional. Esta decisión permite enfocarse específicamente en esta enfermedad y obtener una comprensión más profunda de los datos relacionados.

El análisis exploratorio de datos comienza examinando la distribución de los datos, identificando posibles valores atípicos y analizando la correlación entre variables. También se utilizan gráficos descriptivos, como histogramas y diagramas de caja, para visualizar patrones y tendencias en los datos.

Es importante tener en cuenta que algunos datos clínicos pueden presentar valores nulos debido a que no todos los pacientes se han sometido a todos los exámenes. Por lo tanto, se realiza un análisis cuidadoso y se tiene en cuenta esta limitación al interpretar los resultados.

Se analiza la variable de tratamiento y se encuentra que la mayoría de los pacientes han recibido quimioterapia, cirugías, radioterapia e inmunoterapia, lo cual es coherente con los tratamientos estándar para el cáncer de mama.

El análisis de la variable de estadio del cáncer revela que el estadio II es el más común en la muestra, seguido del estadio I y el estadio III. El estadio IV, que indica un cáncer avanzado que se ha diseminado a otras partes del cuerpo, representa una proporción menor de la muestra. Es importante mencionar que se observa una presencia significativa de valores nulos en esta variable, lo cual puede ser un desafío al desarrollar el modelo predictivo.

Se examina también la variable de metástasis, que indica si hay evidencia de diseminación del cáncer a otras partes del cuerpo. La mayoría de los pacientes muestran una clasificación de M0, lo que sugiere que no hay evidencia de metástasis a distancia en el momento del diagnóstico. Por otro lado, M1 indica que el cáncer se ha diseminado a órganos distantes. Se observa que aproximadamente el 30% de los pacientes no tienen información registrada en este examen.

Se evalúa la clasificación de la afectación de los ganglios linfáticos regionales (PN) y se encuentra que la mayoría de la muestra tiene una clasificación N0, lo que indica que no se encontró evidencia de metástasis en los ganglios linfáticos regionales. La clasificación N1, que indica la presencia de metástasis en los ganglios linfáticos axilares del mismo lado del cuerpo que el tumor primario, sigue a N0 en términos de frecuencia.

Se analiza la variable de tamaño y extensión del tumor primario (PT y CT, respectivamente). Se encuentra que el tamaño T1 es el más común, seguido de T2. Estas clasificaciones indican el tamaño y la extensión del tumor primario, y pueden proporcionar información relevante para el pronóstico y el tratamiento.

En cuanto a la variable de edad, se observa que se presenta información completa sin valores nulos para cada fila. Se identifica una moda alrededor de los 57 años y una concentración de casos entre los 48 y los 66 años. La muestra incluye pacientes desde los 25 años hasta los 97 años. Para facilitar el análisis, se sugiere recodificar la variable de edad en grupos etarios, lo cual permitirá una mejor comprensión de las características de los pacientes en función de su edad.

En resumen, este análisis exploratorio proporciona una visión general y descriptiva de los datos relacionados con el cáncer de mama. Se destacan las características de la muestra, como la distribución geográfica, los tratamientos recibidos y los estadios del cáncer. Se resalta la presencia de valores nulos en algunas variables y se sugiere la recodificación de la variable de edad para facilitar el análisis. Estos hallazgos sientan las bases para el posterior desarrollo del modelo predictivo utilizando técnicas de Machine Learning.

# Preprocesamiento de los datos

Para el preprocesamiento de la base de datos, se realizaron diversas acciones:

* Se generó una copia de la base de datos original para mantener los datos intactos.
* Se filtraron los datos seleccionando únicamente las filas relacionadas con el cáncer de mama.
* Se recodificaron las variables de fecha a un formato datetime y se calculó la duración de los tratamientos y la espera entre el diagnóstico y el inicio del tratamiento.
* Se agrupó la variable de edad en rangos etarios establecidos por el INE.
* Se creó la columna "STATUS" como el vector objetivo utilizando la variable "ESTADIO", asignando valores numéricos según la clasificación del estadio del cáncer.
* Se transformaron las variables categóricas restantes en variables indicadoras (dummies).
* Se eliminaron las filas que tenían valores faltantes en la variable "ESTADIO".
* Se reemplazaron los valores NaN con "No Determinado" en las filas que tenían al menos un valor distinto a NaN en las columnas relacionadas con el detalle del tumor.
* Estos pasos de preprocesamiento permitieron tener una base de datos lista para el modelamiento y análisis posterior, con variables adecuadamente codificadas y sin valores faltantes en la variable objetivo.

# Análisis exploratorio post procesamiento

Tras el procesamiento de los datos, se lleva a cabo un nuevo análisis exploratorio para evaluar el impacto de los cambios y realizar análisis multivariados con los datos procesados.

En cuanto a la variable "ESTADIO", se ha reclasificado y se observa que la mayoría de la muestra pertenece al "STATUS 1" (39.04%), indicando un cáncer en etapa inicial. El "STATUS 2" representa el 36.24% de los pacientes, correspondiente a un cáncer en etapa intermedia, mientras que el "STATUS 3" (cáncer avanzado) constituye el 24.7% de la muestra.

En las variables procesadas, los valores nulos ahora se denominan "NO DEFINIDO", siguiendo la misma tendencia que antes del procesamiento. El gráfico muestra un cambio en la distribución de los beneficiarios del convenio oncológico, con un predominio de pacientes no beneficiarios de FALP.

En cuanto al rango etario, la muestra se divide principalmente entre "Adultos" (51.37%) y "Adultos Mayores" (48.12%), con solo 7 jóvenes presentes en los datos.

El análisis de las variables de convenio oncológico y rango etario en relación con la probabilidad de tener un "STATUS" alto revela que los pacientes con ISAPRE tienen una menor probabilidad de tener un cáncer avanzado. En contraste, aquellos con FONASA, particular o beneficiario y sin convenio oncológico tienden a tener una mayor probabilidad de tener un cáncer avanzado.

Al examinar estas variables en relación con la probabilidad de tener un "STATUS" alto, se destaca que los jóvenes tienen una alta probabilidad de tener un "STATUS" alto con un diagnóstico local o regional, mientras que la probabilidad de tener un cáncer avanzado es nula. En el caso de los adultos y adultos mayores, existe una mayor tendencia a tener un "STATUS" alto cuando el cáncer se ha diseminado en otras partes del cuerpo. Se destaca que cuando el cáncer se encuentra solo en la mama (etapa local), la probabilidad de tener un "STATUS" alto es baja.

Estos hallazgos resaltan las diferencias en la distribución del cáncer según el convenio oncológico, la edad y los estadios del cáncer, proporcionando información relevante para comprender la relación entre estas variables y el pronóstico de la enfermedad.

# Modelación y predicción del proyecto

El problema abordado es una clasificación multiclase supervisada para predecir la urgencia de atención de los pacientes según variables clínicas, patológicas y sociodemográficas. Se entrenan varios modelos de clasificación multiclase, como Naive Bayes Multinomial, Árbol de Clasificación Simple, Random Forest de Clasificación, Gradient Boosting de Clasificación y Support Vector Machine. Se realizan búsquedas de hiperparámetros para encontrar los modelos más óptimos.

El conjunto de datos preprocesado se divide en 66.7% para entrenamiento y 33.3% para evaluación del desempeño. Se utilizan métricas como exactitud, recall y precisión para evaluar el desempeño de los modelos.

Para cada modelo, se exploran diferentes valores de hiperparámetros. Por ejemplo, para Naive Bayes Multinomial se ajusta el parámetro alpha y fit\_prior, mientras que para Árbol de Clasificación Simple se ajustan max\_depth, max\_leaf\_nodes, criterion y min\_samples\_split. Se realizan ajustes similares en los demás modelos.

Después de entrenar y evaluar los modelos, se obtienen resultados que indican el desempeño de cada modelo en términos de métricas. En general, la mayoría de los modelos presentan un buen desempeño con valores superiores a 0.9, excepto Naive Bayes. El modelo Gradient Boosting muestra el mejor desempeño en términos de recall en las clases 1 y 2, y en accuracy de validación. Por lo tanto, se decide implementar el modelo Gradient Boosting Classifier debido a su mayor acierto en la variable STATUS 3, que corresponde a los pacientes de mayor gravedad.

# Proyección y solución

Con el fin de utilizar la investigación y predicción realizada en este proyecto, se ha desarrollado una demo en forma de formulario web. Este formulario permite ingresar los datos de un paciente, incluyendo los resultados de sus exámenes clínicos o patológicos. Se han establecido requisitos para el ingreso de datos con el objetivo de asegurar una entrada de datos limpia y mejorar los resultados del modelo predictivo.

El formulario web incluye campos como EDAD, PREVISION, REGION, COMUNA, EXTENSION\_DIAGNOSTICO, TTO\_FALP\_SUBCATEGORIA, así como las descripciones TNM patológicas y clínicas. Se han implementado controles de entrada, como TextBox, DropDown, RadioButton y CheckBox, para asegurar la selección correcta de valores y evitar errores tipográficos. Además, se ha establecido que al menos uno de los campos de descripción clínica o patológica del tumor debe ser completado.

La aplicación web fue desarrollada utilizando la herramienta Anvil y un notebook en Google Colab, lo que permite que funcione como su propio servidor para montar la demo. Una vez que los datos son ingresados y se presiona el botón "Ingresar", estos datos son procesados y pasan por el modelo predictivo. La predicción resultante se muestra en la aplicación, la cual verifica si el valor de la predicción STATUS es igual a 3. Si se cumple esta condición, la aplicación muestra una alarma que informa al usuario que el paciente recién ingresado tiene un ESTADIO III o IV y debe recibir atención médica de manera urgente.

El propósito de esta aplicación web es servir como una herramienta de priorización en la atención de pacientes, ayudando en la identificación de casos de cáncer que requieren una atención inmediata. Esto permite colaborar con los médicos oncólogos en la declaración de la gravedad del cáncer en los paciente.

# Conclusiones

Los resultados obtenidos en la implementación de modelos de machine learning demuestran que son una poderosa herramienta para predecir el comportamiento del cáncer de mama en los pacientes. Mediante un modelo predictivo de clasificación, se logra identificar patrones y realizar predicciones de estadio con una precisión superior al 95%.

Es fundamental contar con un registro adecuado de los datos clínicos y patológicos del paciente para lograr una correcta predicción de estadio. Por lo tanto, se propone el uso de una aplicación web que permita ingresar y estructurar los datos de manera ordenada. Esto facilitará el trabajo con una mayor cantidad de datos y reducirá la presencia de valores nulos o vacíos.

La predicción del estadio mediante esta herramienta alertará a los gestores de casos oncológicos a través de una alarma, indicando cuándo un paciente se encuentra en situación de riesgo. Esto permitirá a los profesionales clínicos priorizar la atención, detectar el cáncer de manera temprana y precisa, y brindar un tratamiento adecuado. Asimismo, fomentará la atención preventiva en salud.

En conclusión, una herramienta de machine learning junto con una aplicación web para la estandarización y disponibilidad de datos clínicos pueden predecir el estadio del cáncer de mama con una precisión del 95%. Esto se traduce en un apoyo confiable para los profesionales de la salud, generando datos de calidad y facilitando la clasificación del estadio. Esta herramienta puede ser implementada en diferentes niveles de atención pública y privada, y representa un avance en el campo de la salud digital.