

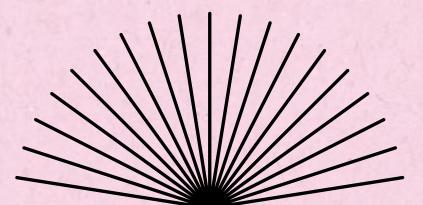


The Bridge
21/02/2025

MACHINE LEARNING/Project

**Predicción del Overall Survival y Relapse Free Survival
de Pacientes con Cáncer de Mama**

DANIEL MASANA DIEGO



Descripción del Data set

Descripción datos:

2509 pacientes

34 columnas/variables

float64(10), object(24)

Origen dataset

METABRIC (Molecular Taxonomy of Breast Cancer International Consortium):

- Contiene información clínica y genómica de más de 2500 pacientes,
- recurso valioso para el desarrollo de modelos de predicción y clasificación de subtipos tumorales.



Datos faltantes

- Varias columnas tienen una cantidad significativa de valores nulos, lo que requerirá estrategias de imputación o eliminación.

Variables Mixtas

- Combinación de datos numéricos (ej.: edad, tamaño de tumor) y categóricos (ej.: tipo de cancer, estado hormonal).

Potencial para Análisis

- Las columnas relacionadas con supervivencia y tratamiento permitirán análisis predictivos o de correlación.

Variables Categóricas

Procedimiento Quirúrgico:

- Type of Breast Surgery: Mastectomy/Breast conserving

Estado hormonal:

- ER Status (+,-)
- PR Status (+,-)
- HER2 Status (+,-)

Características Clínicas:

- Primary Tumor Laterality (Left/Right)
- Inferred Menopausal State (Pre/Post)

Tratamientos:

- Chemotherapy (Yes/No)
- Hormone Therapy (Yes/No)
- Radio Therapy (Yes/No)

Estado de Supervivencia:

- Overall Survival Status (Living/Deceased)

Variables Numéricas

Datos clínicos y demográficos

- Age at Diagnosis
- Cohort

Características tumorales:

- Tumor Size
- Tumor Stage
- Lymph nodes examined positive
- Mutation Count
- Nottingham prognostic index
- Neoplasm Histologic Grade

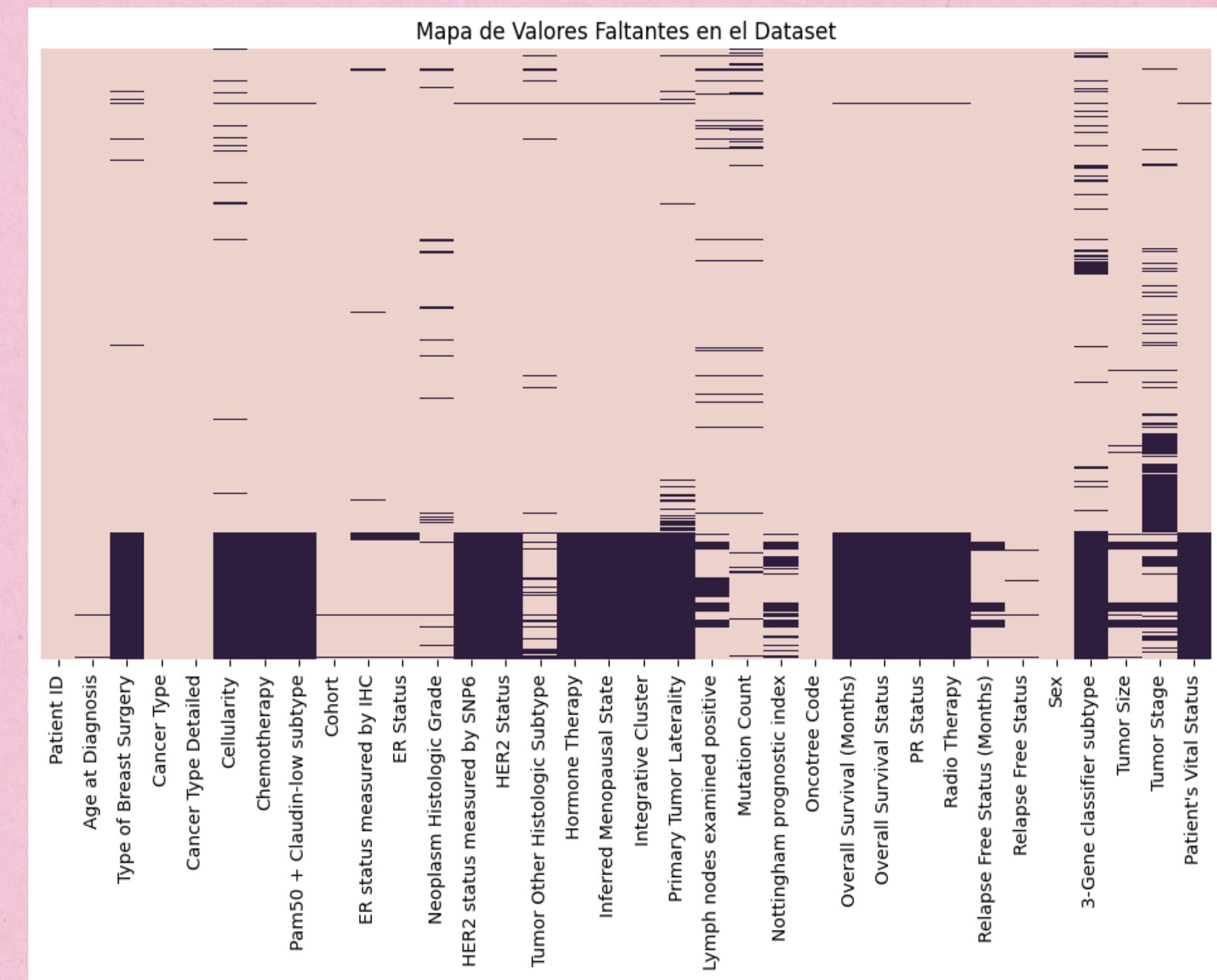
Supervivencia y recaída:

- Overall Survival Grade (Months)
- Relapse Free Status (Months)

Valores faltantes y eliminación

Eliminación Antes de la Imputación:

- Patient ID
- Cohort
- Sex



Estrategias de imputación

Categoría	Variable	Método de Imputación
Variables de Tratamiento	"Chemotherapy", "Hormone Therapy", "Radio Therapy"	Moda dentro de cada Cancer Type Detailed.
Variables Relacionadas con el Cáncer	"Type of Breast Surgery"	Moda basada en Cancer Type Detailed.
	"Cellularity"	Moda dentro de Cancer Type Detailed.
	"Neoplasm Histologic Grade"	Mediana dentro de Tumor Stage.
	"Tumor Other Histologic Subtype"	Moda dentro de Cancer Type Detailed.
Variables Moleculares y Biomarcadores	"Pam50 + Claudin-low subtype", "HER2 Status", "PR Status", "ER Status"	Moda en función de Cancer Type Detailed y otros biomarcadores.
	"HER2 status measured by SNP6"	Basado en Her2S Status
	"ER status measured by IHC"	Moda basada en ER Status.
Variables Clínicas y Demográficas	"Inferred Menopausal State"	Basado en ER Status, PR Status y Age at Diagnosis.
	"Patient's Vital Status"	Asignar basado en Overall Survival Status.
Variables con Asignación Directa	"3-Gene classifier subtype"	Asignar Unknown a los valores nulos.
	"Integrative Cluster"	Asignar Unknown a los valores nulos.

Estrategias de imputación

Variable	Método de Imputación	Justificación
Age at Diagnosis	Mediana por subtipo de cáncer	La edad de diagnóstico varía por tipo de tumor.
Lymph nodes examined positive	Regresión lineal con Nottingham prognostic index	Están fuertemente correlacionados.
Mutation Count	Mediana	No tiene correlaciones fuertes.
Nottingham prognostic index	Regresión con Lymph nodes examined positive	Relación fuerte entre estas variables.
Overall Survival (Months)	Regresión con Relapse Free Status (Months)	Alta correlación (~0.8).
Relapse Free Status (Months)	Regresión con Overall Survival (Months)	Alta correlación.
Tumor Size	Mediana por etapa tumoral y subtipo de cáncer	Tamaño tumoral varía según estadio y tipo molecular.
Tumor Stage	Clasificación basada en Tumor Size	Moderada correlación entre tamaño y estadio.

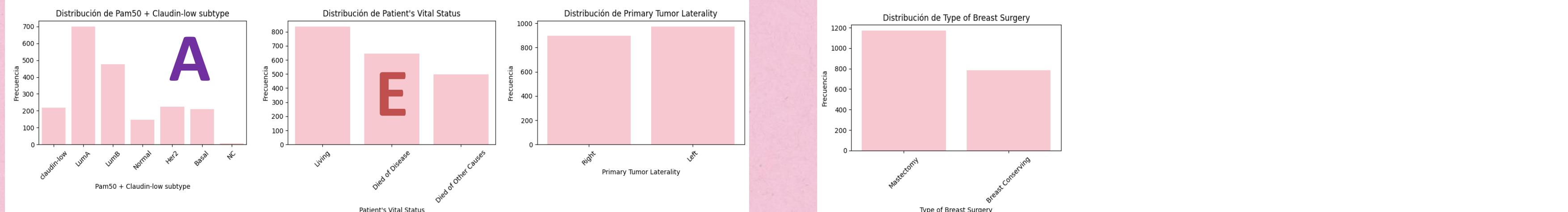
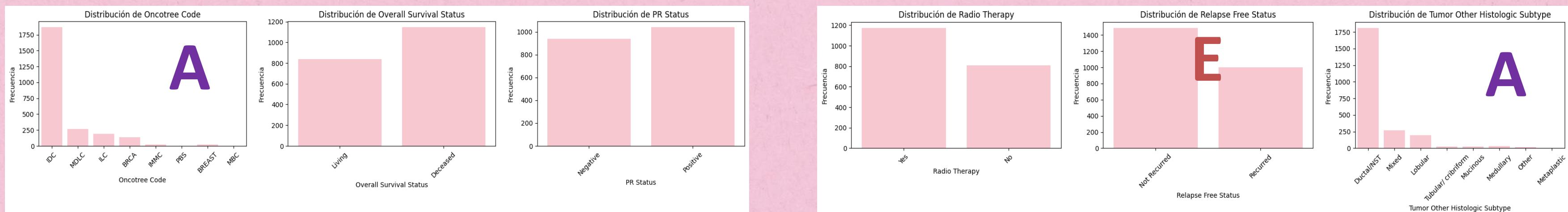
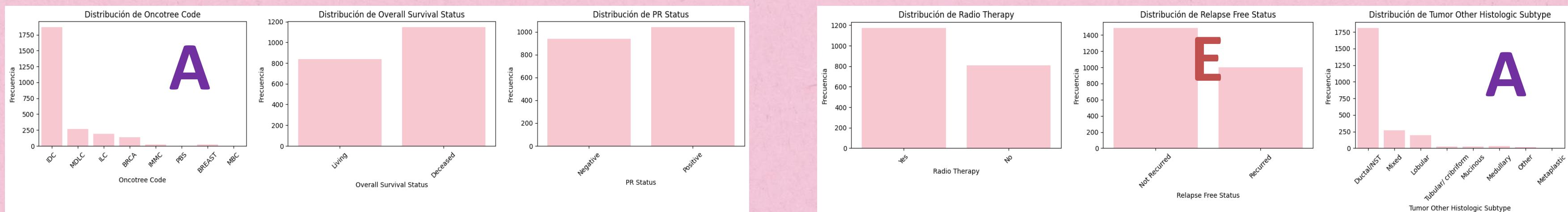
Eliminación y Agrupación después de la Imputación

Eliminación:

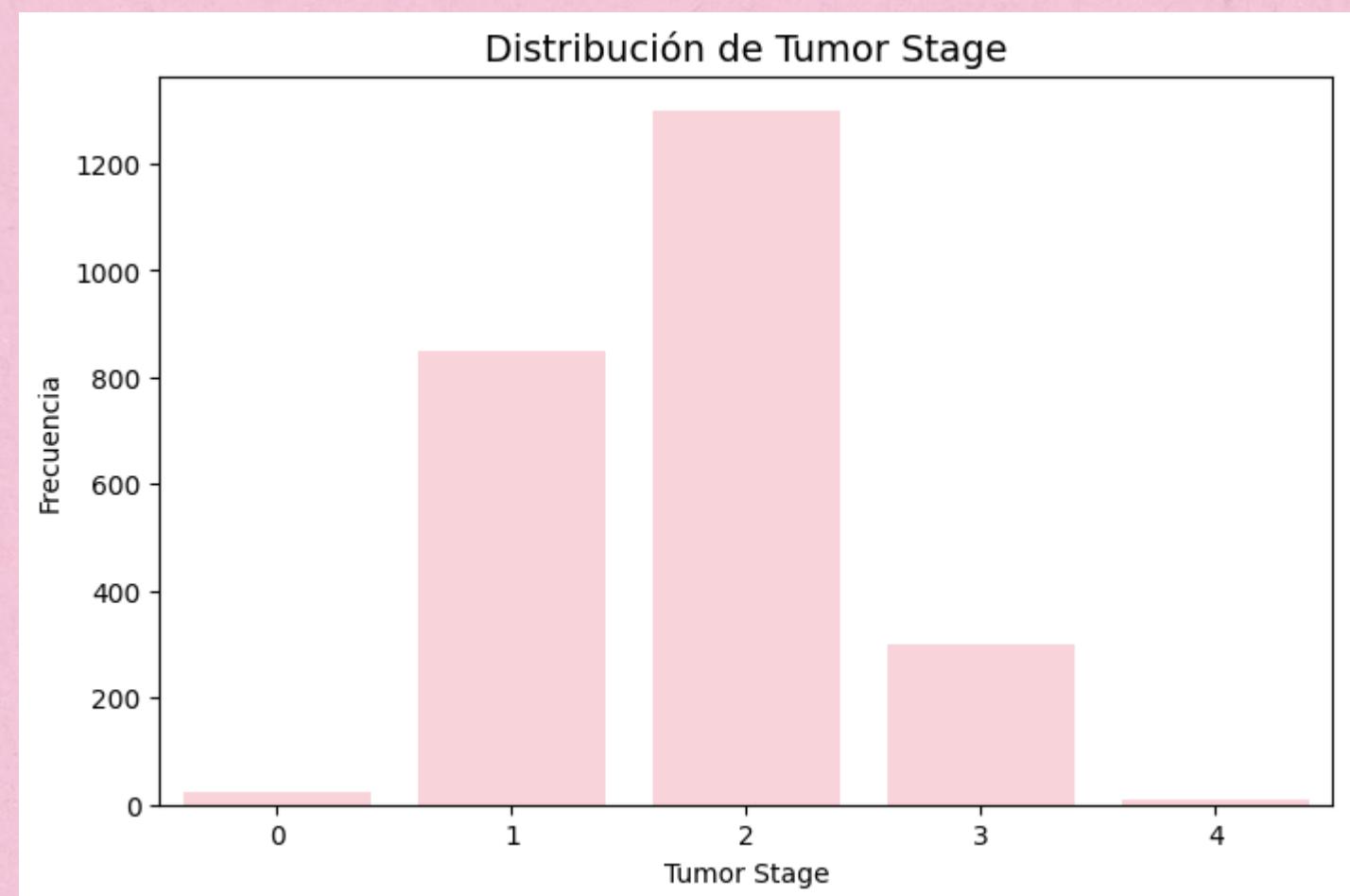
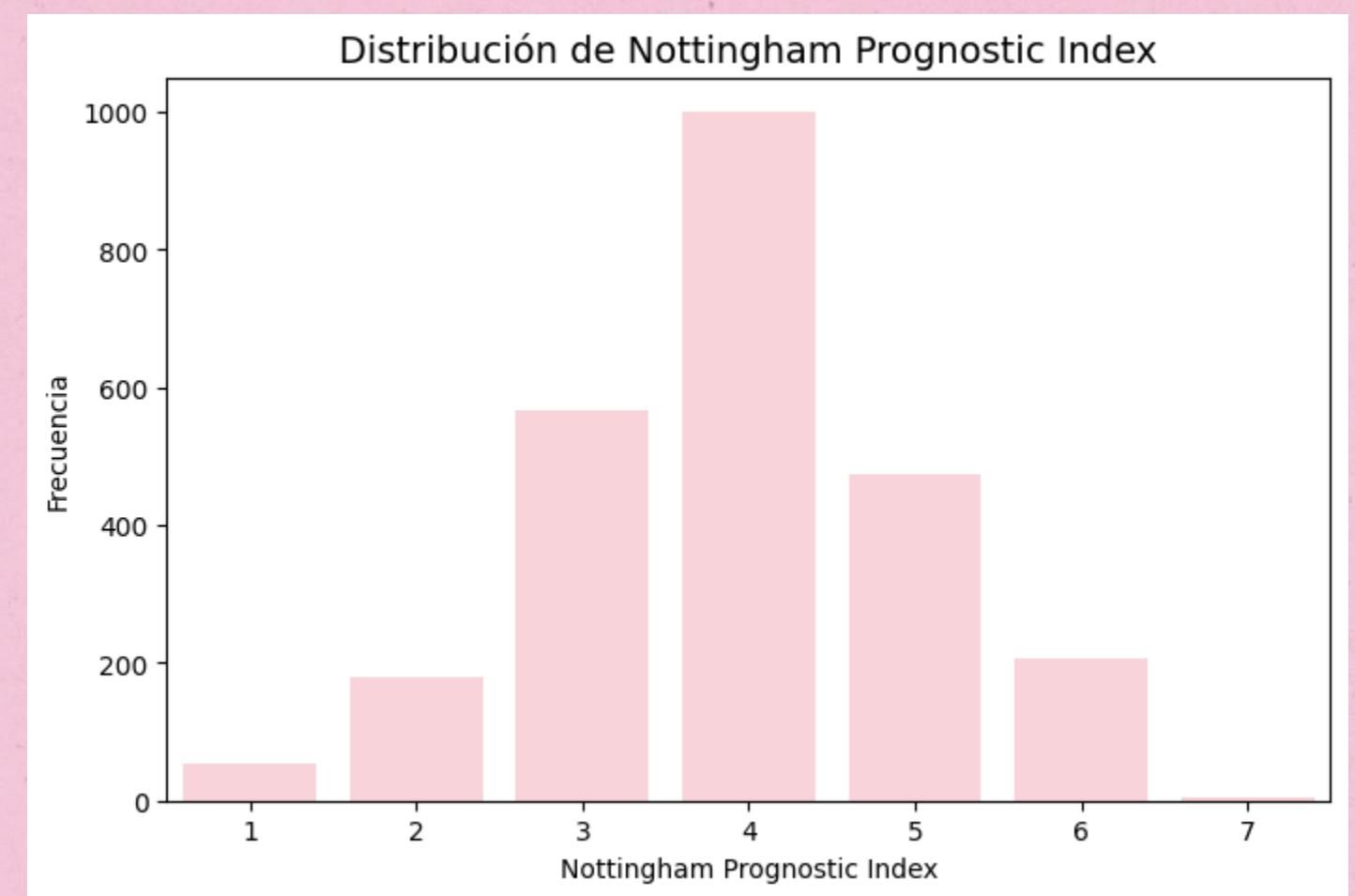
- 3-Gene classifier subtype
- Relapse Free Status
- Integrative Cluster
- Cancer Type Detailed
- HER2 status measured by SNP6

Agrupación:

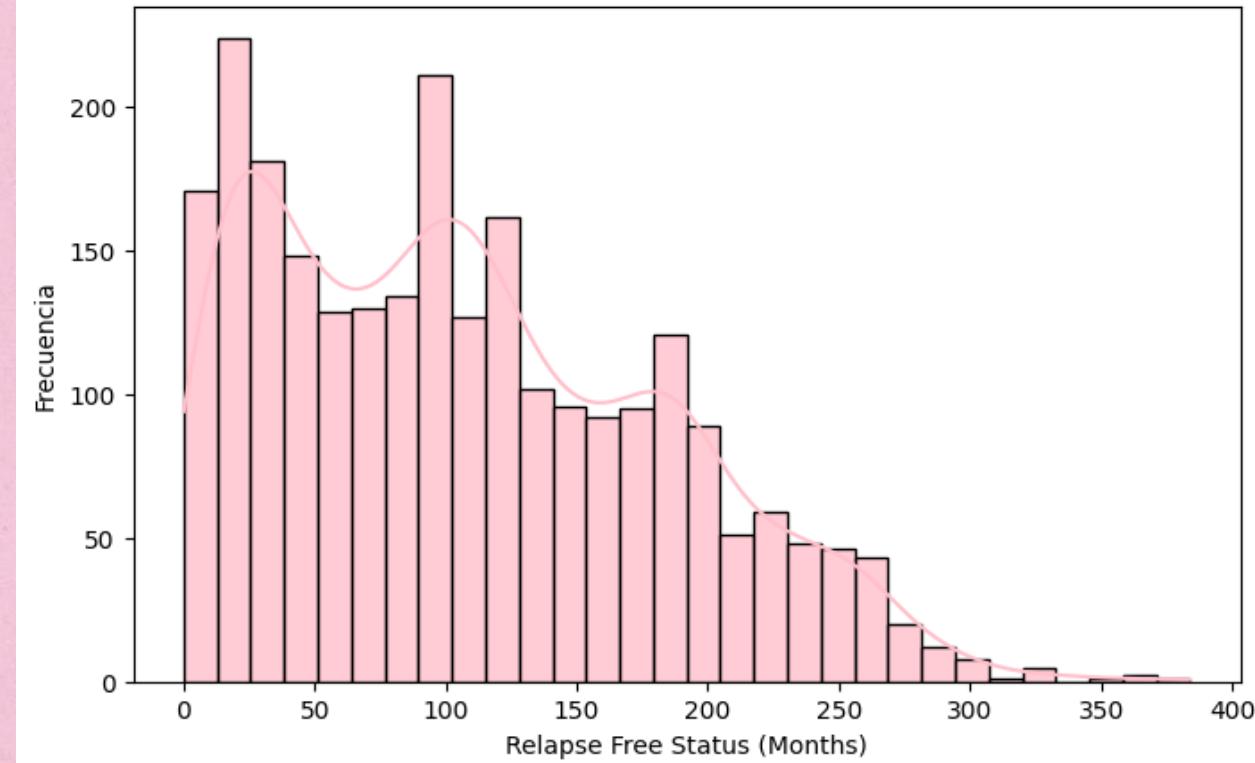
- Cancer Type Detailed
- Pam50 + Claudin-low Subtype
- Tumor Other Histologic Subtype
- Oncotree Code



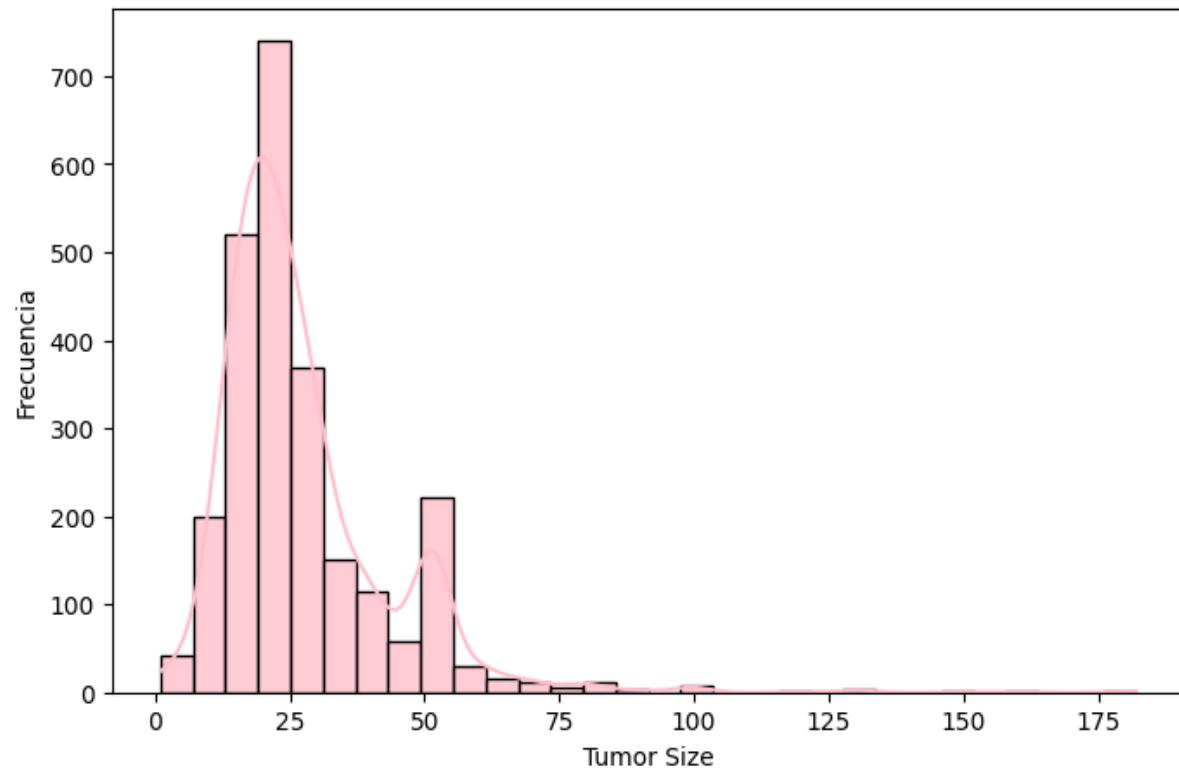
Análisis de la distribución de variables numéricas



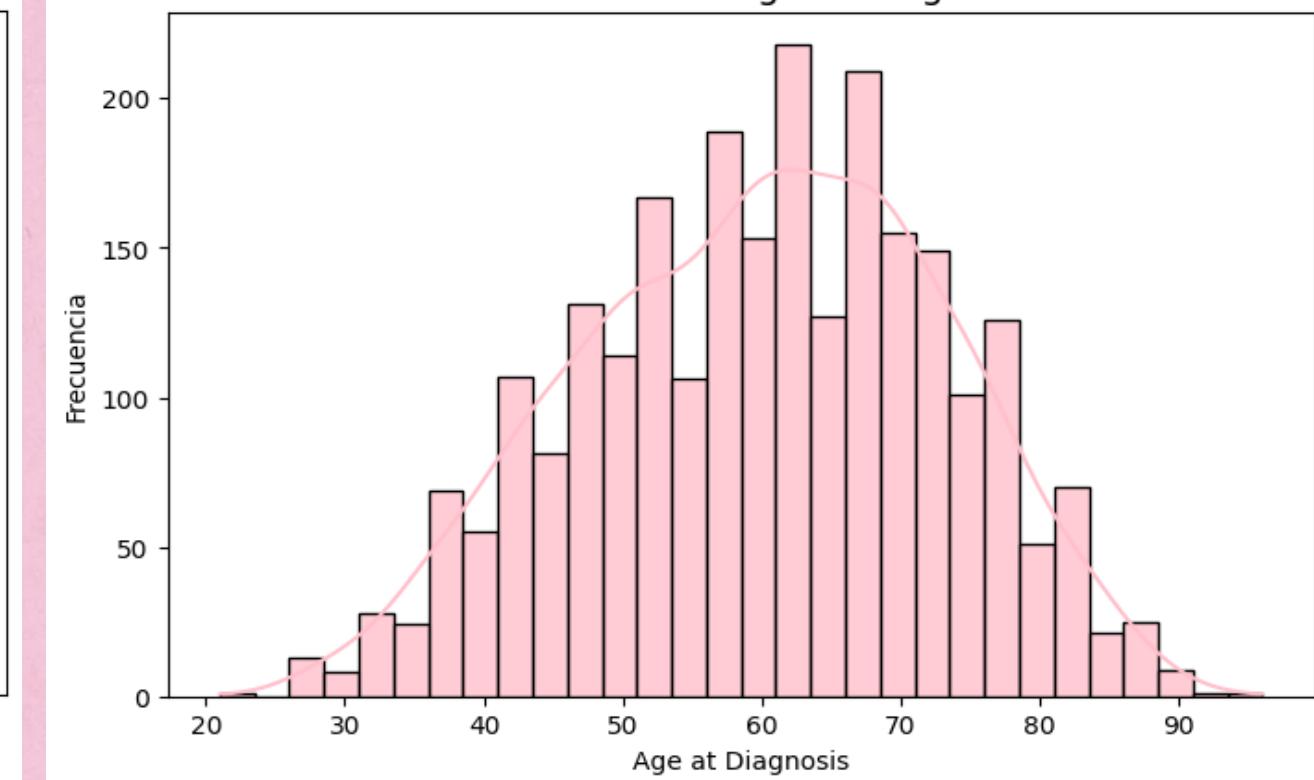
Distribución de Relapse Free Status (Months)



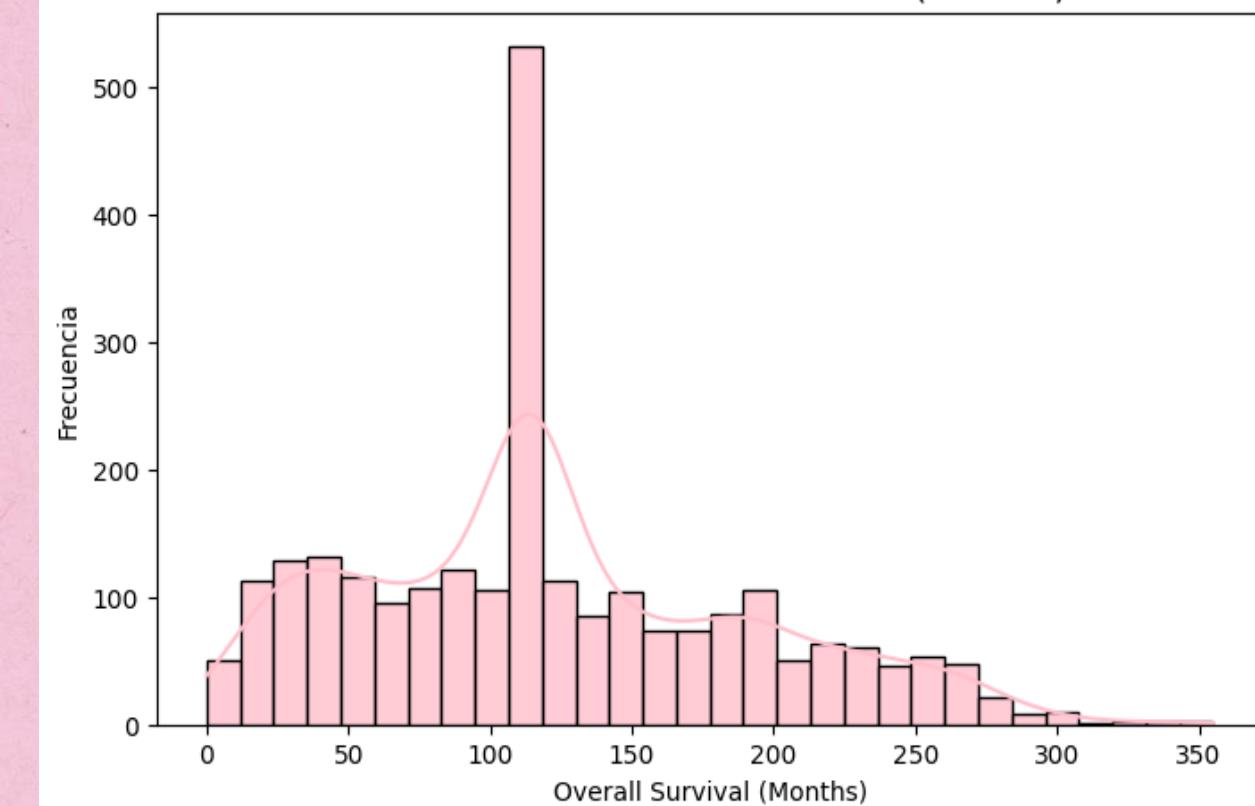
Distribución de Tumor Size



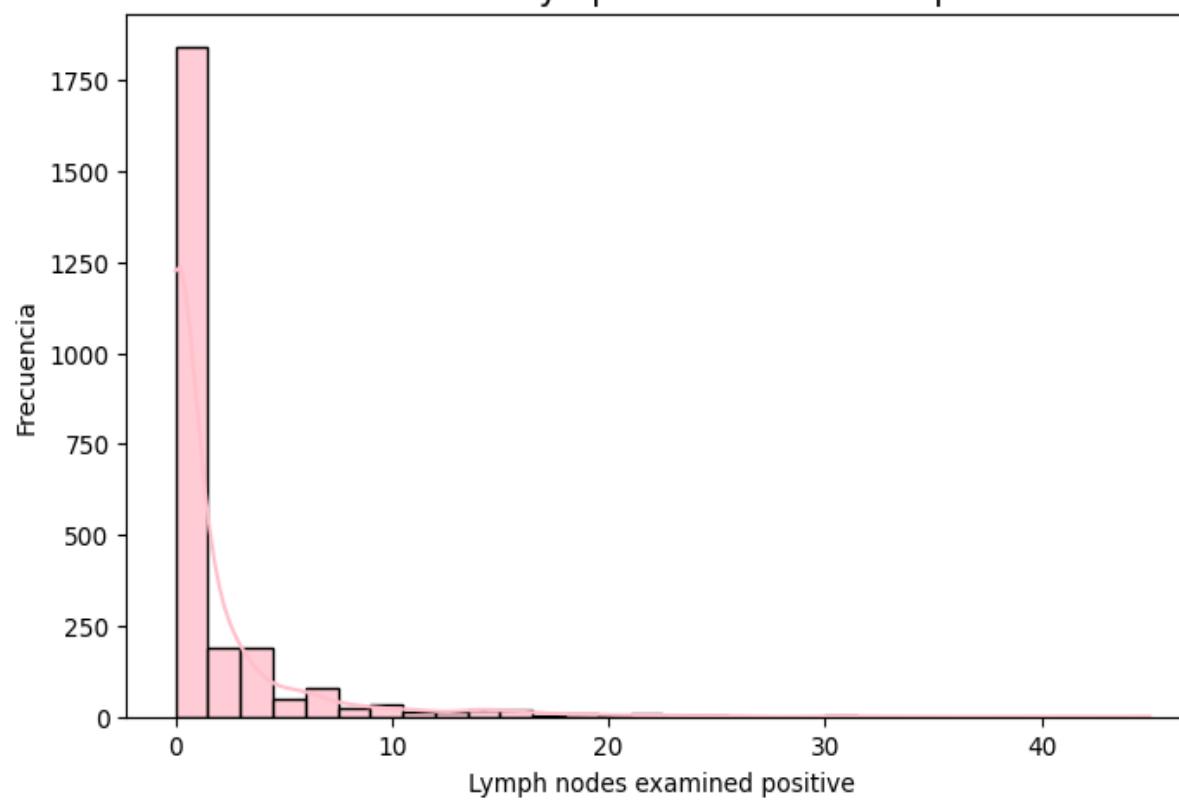
Distribución de Age at Diagnosis



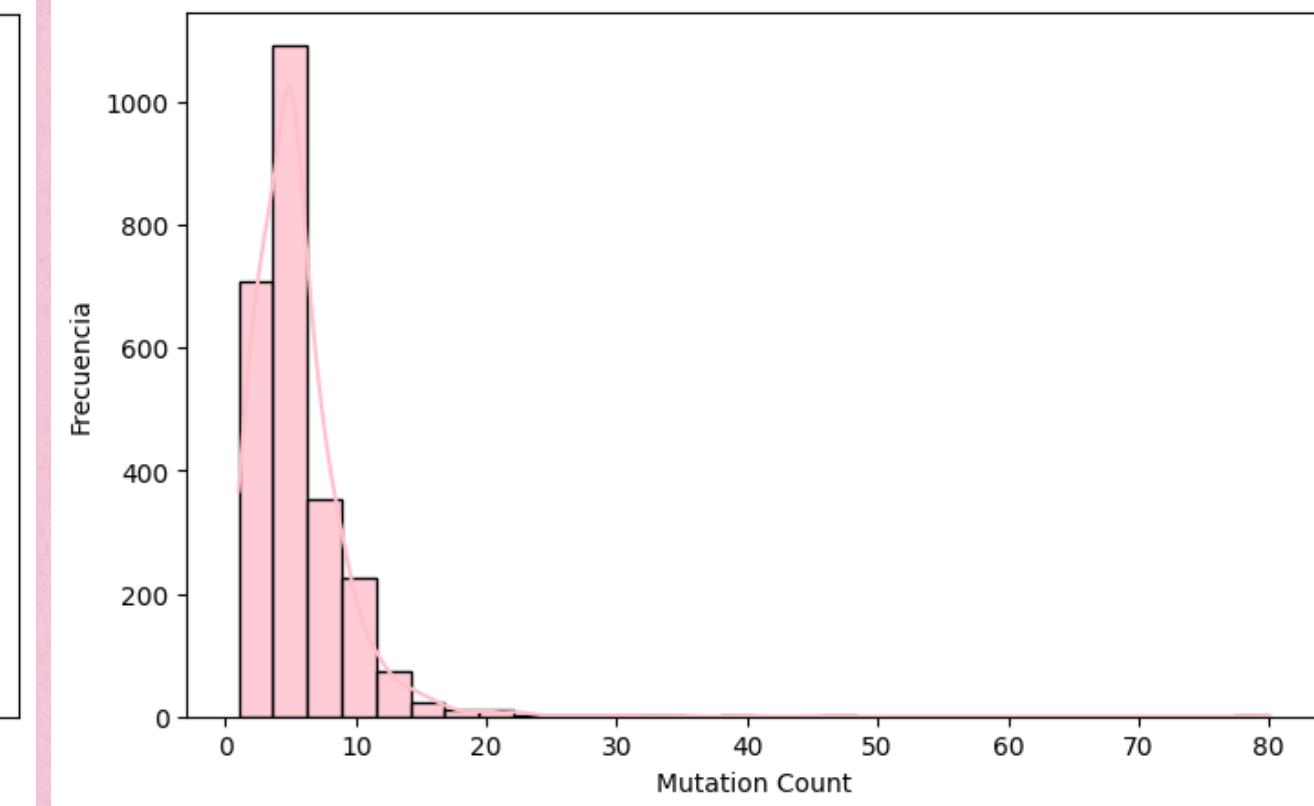
Distribución de Overall Survival (Months)



Distribución de Lymph nodes examined positive



Distribución de Mutation Count



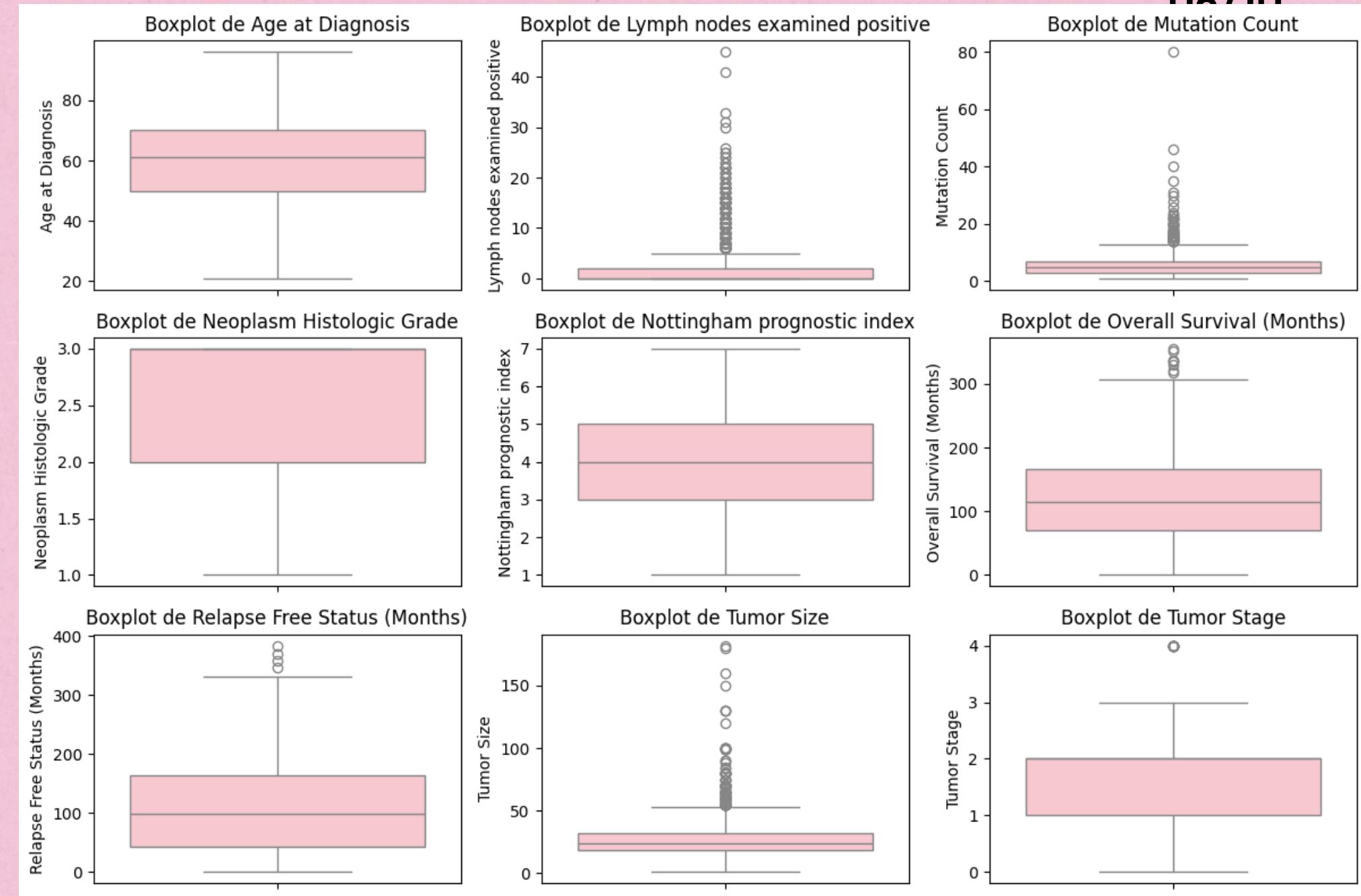
Análisis de Outliers

En variables Categóricas:

No se observaron Missing values importantes.

En variables Numéricicas

Establecimos límites para 3 variables



• Tumores más grandes tienden a estar en etapas más avanzadas.

Evaluación de correlaciones

Relapse Free Status (Months) y Overall Survival (Months) (Fuerte)

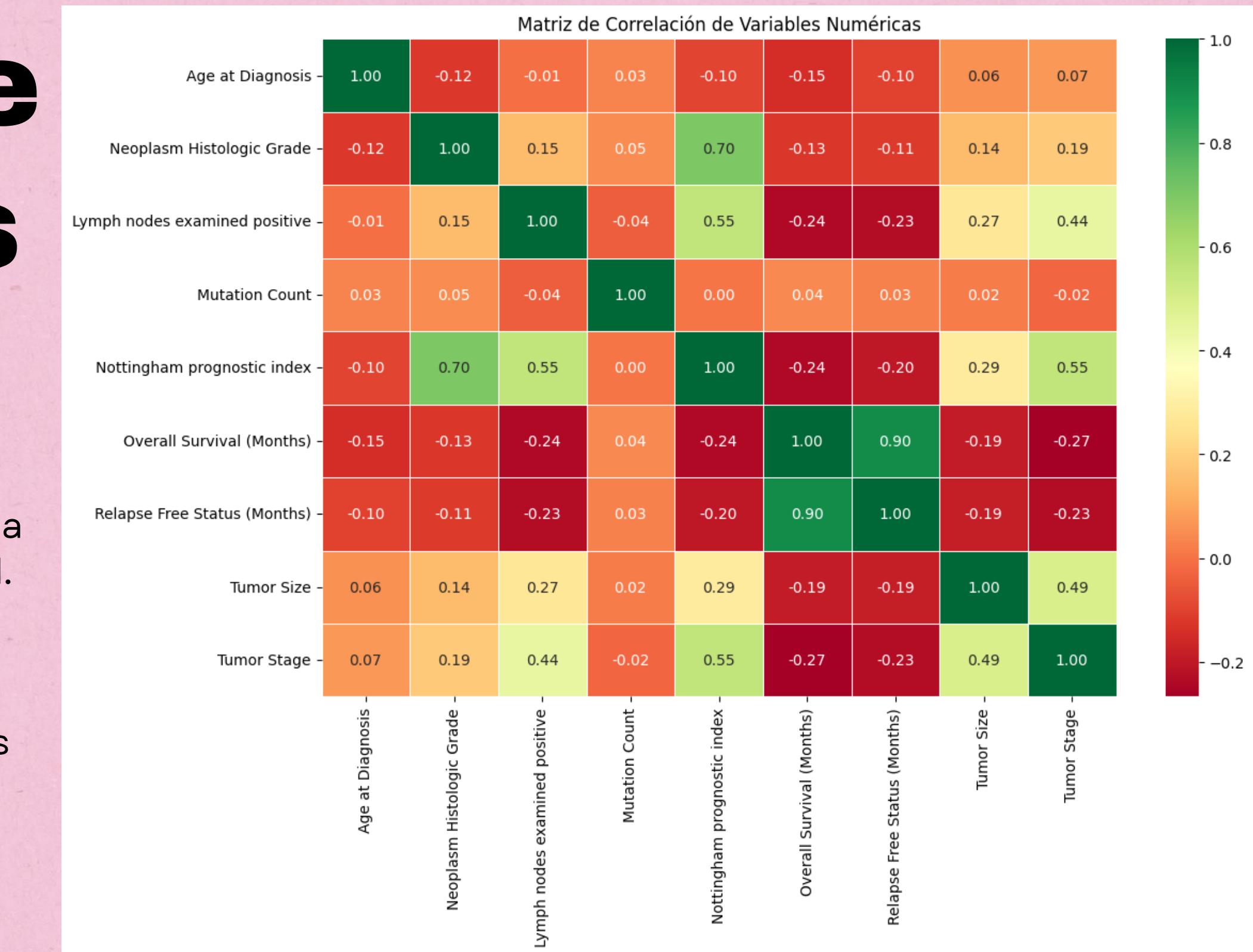
- Pacientes que tienen un mayor tiempo sin recaída tienden a tener una mayor supervivencia general.

Tumor Size y Tumor Stage (Media):

- Tumores más grandes tienden a estar en etapas más avanzadas.

Nottingham Prognostic Index (NPI) con Tumor Size y Tumor Stage (Media):

- Este índice, utilizado como predictor de supervivencia, está altamente relacionado con estas características del tumor.



Feature Engineering

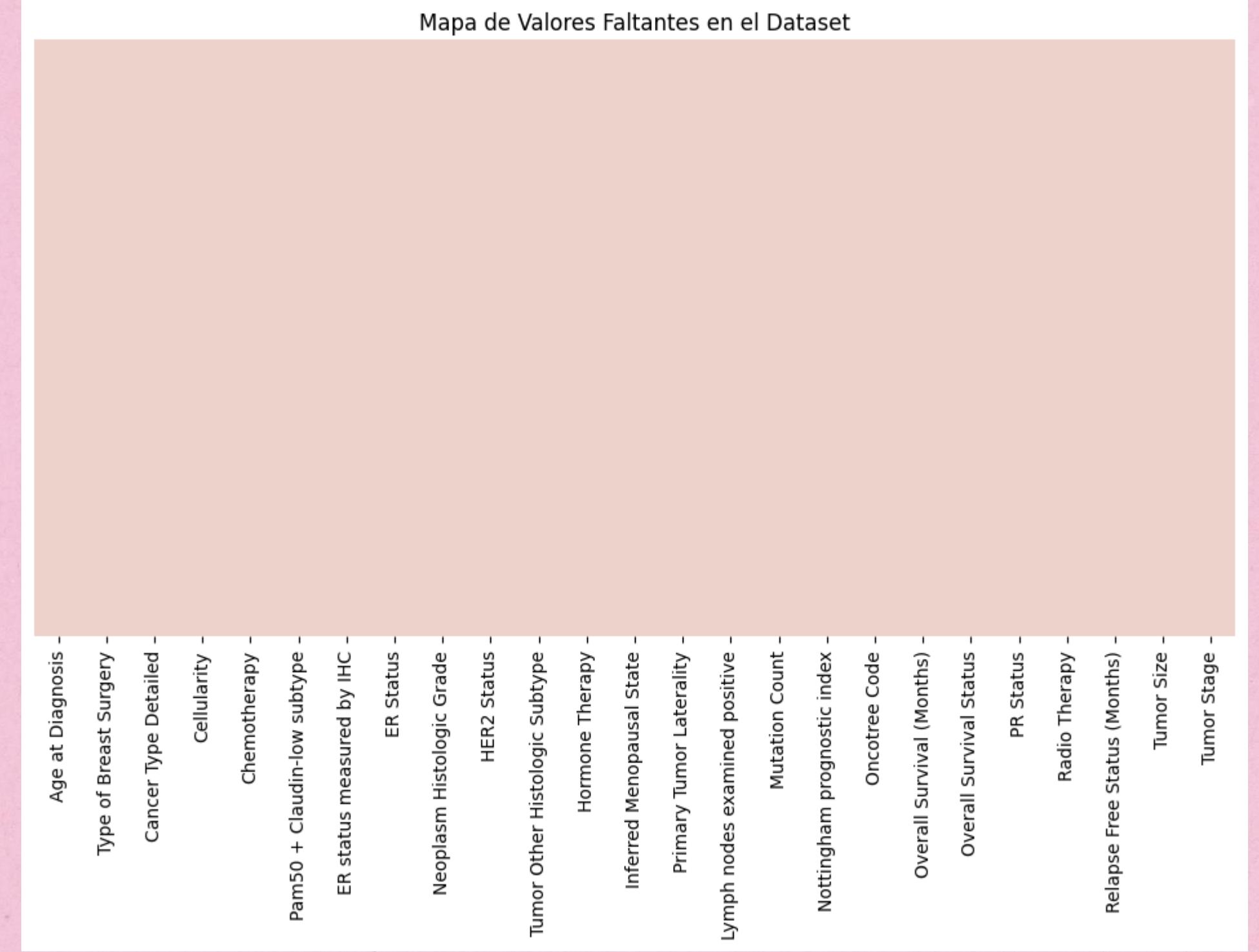
Punto de partida:

- 2484 pacientes de 2509
- De 34 a 25
- 9 Numéricas
- 16 Categóricas

Detección de variables numéricas redundantes:

- Correlation threshold en 0.95
- No hay redundancias
- Neoplasm Histologic Grade (Posible eliminación)

Mapa de Valores Faltantes en el Dataset



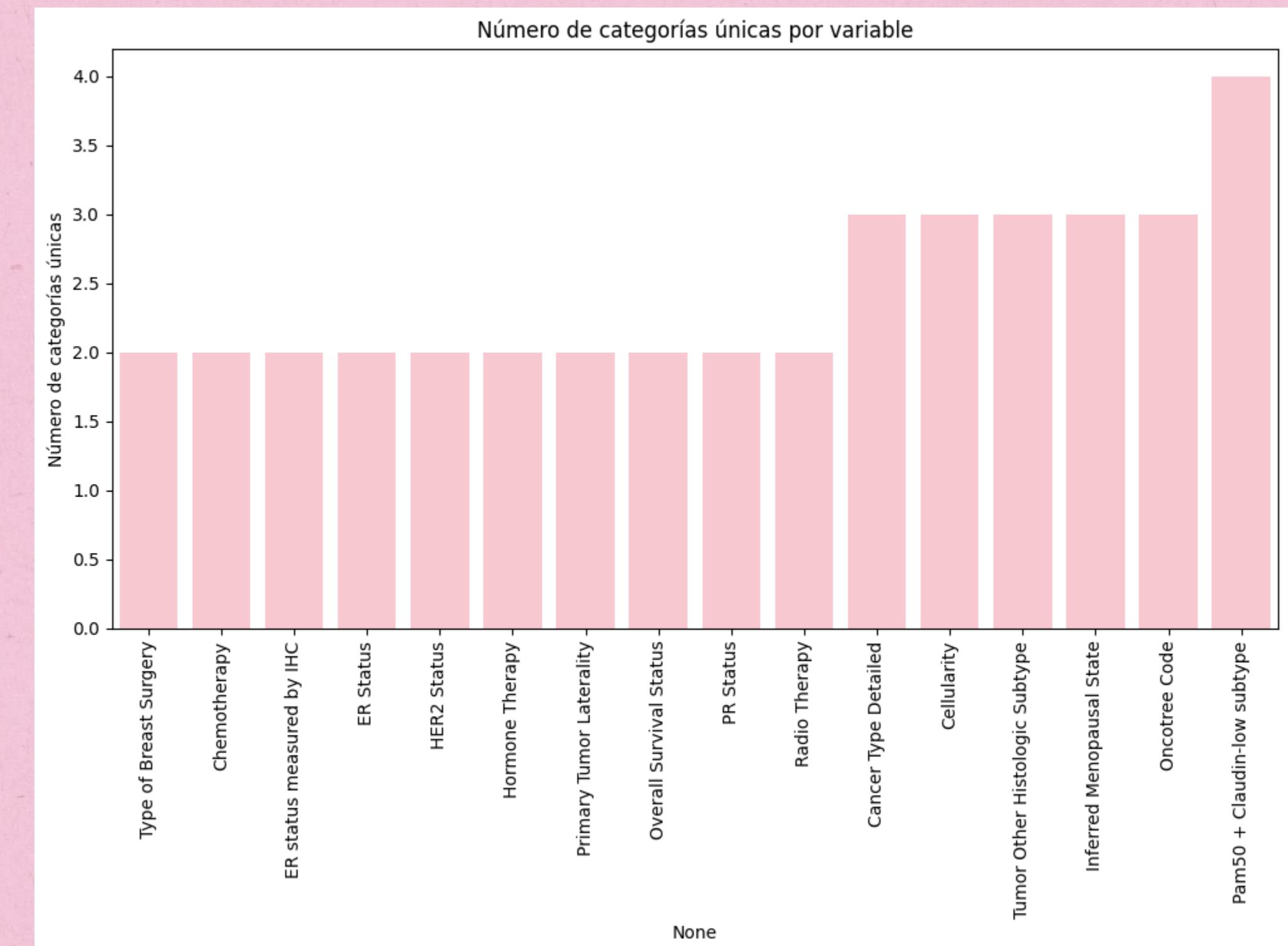
Feature Engineering

Conversión a Binario:

- Visualizamos cuanta categorias unicas hay.
- Pasamos a binario las 11 variables.

Codificación de variables categóricas:

- Previamente agrupadas
- One hot encoding
- Ordinal category

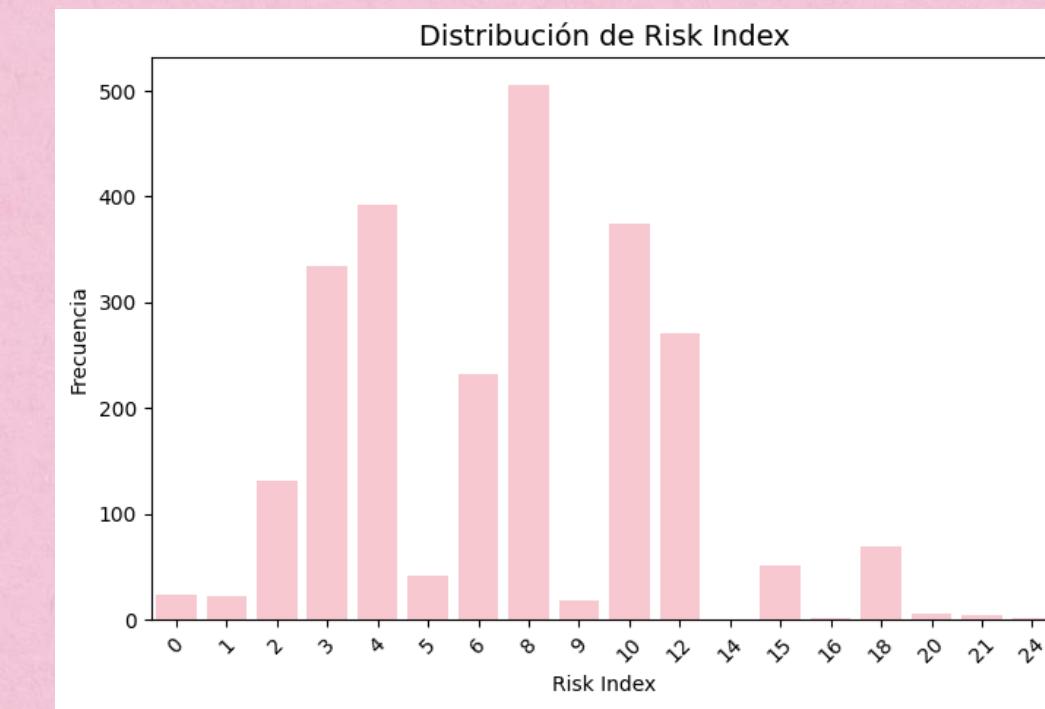
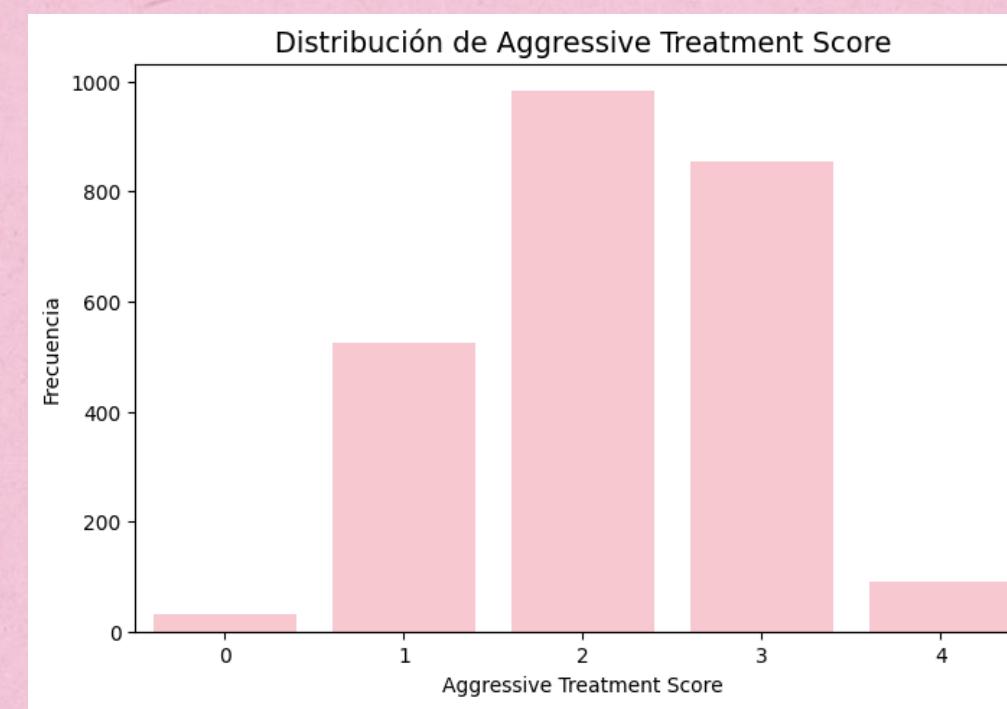
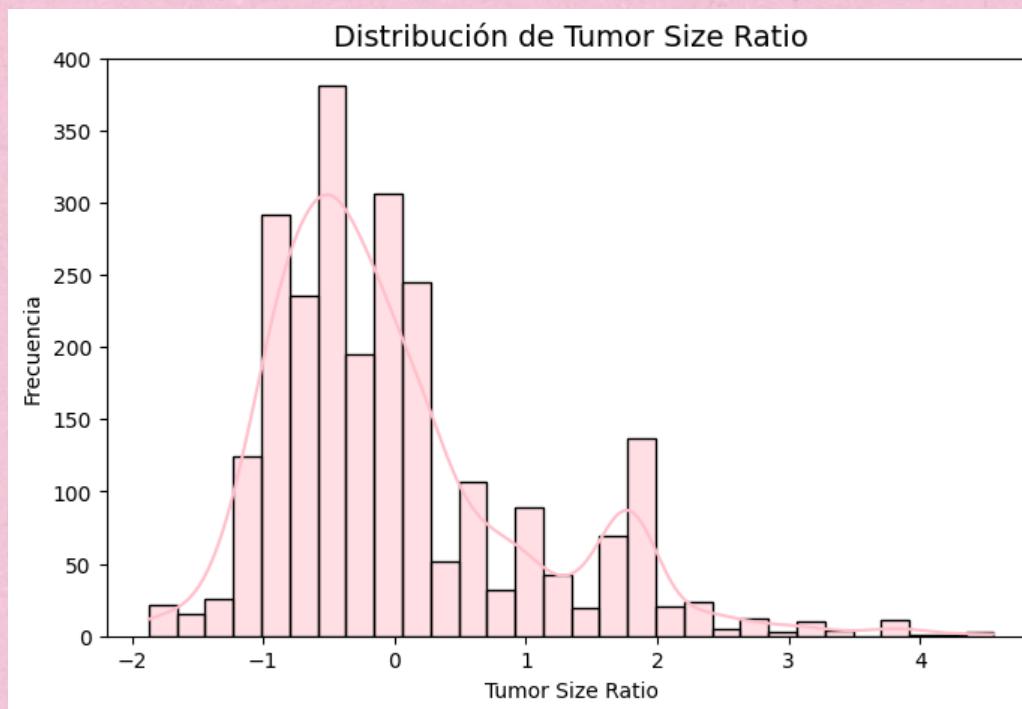


Feature Engineering

Creación de variables Derivadas:

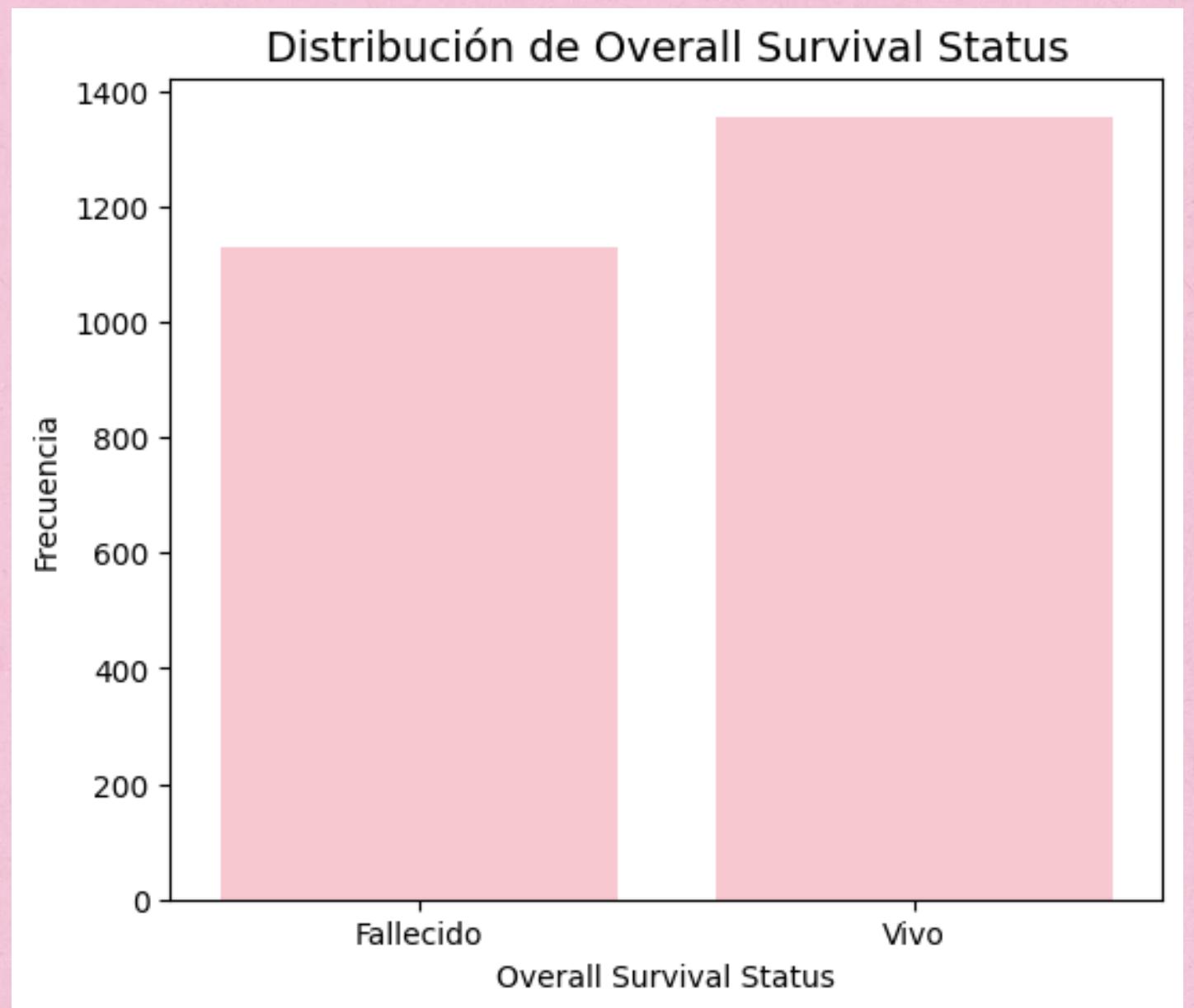
- Tumor Size Ratio
- Risk Index
- Aggressive Index Score

Variable Derivada	Correlación con Overall Survival Status	Correlación con Relapse Free Status (Months)
Aggressive Treatment Score	0.147 (positiva)	-0.150 (negativa)
Survival Ratio	0.018 (casi nula)	-0.203 (moderada-negativa)
Tumor Size Ratio	-0.066 (débil-negativa)	-0.172 (moderada-negativa)
Risk Index	-0.109 (débil-negativa)	-0.221 (moderada-negativa)



Overall Survival Status

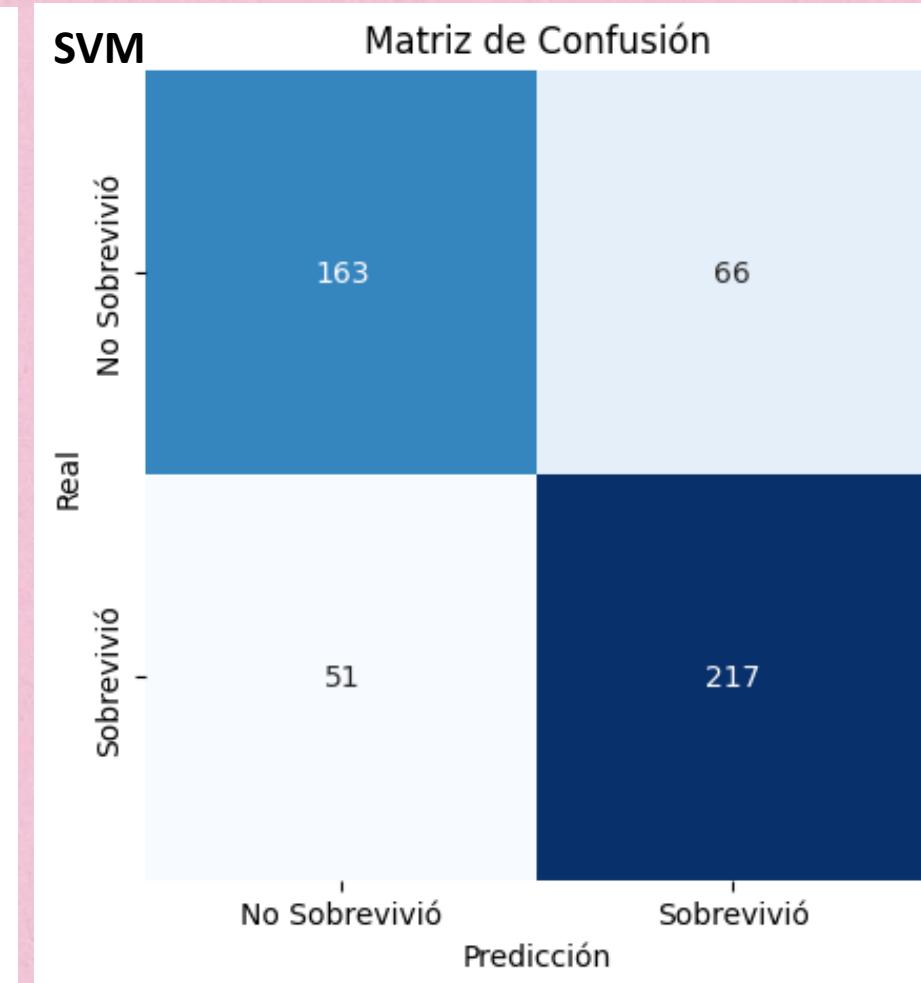
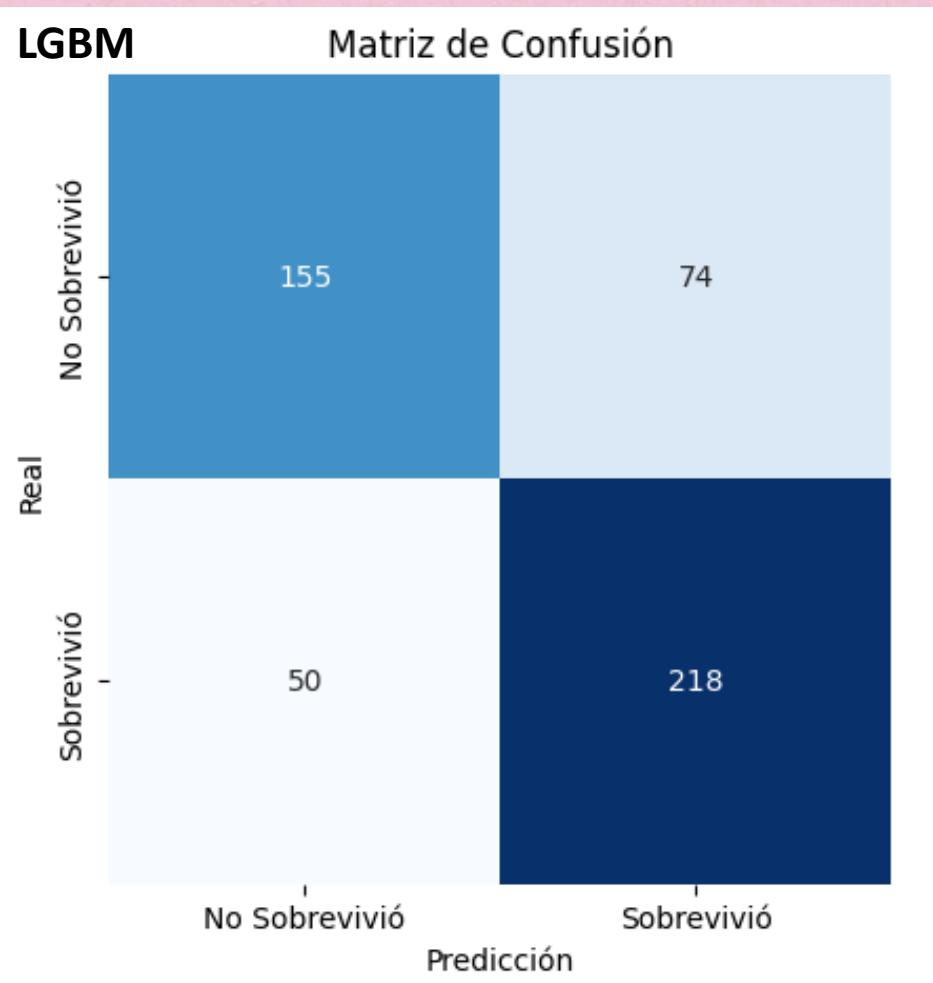
- Aggressive Treatment
- Mutation count
- Nottingham prognostic index score
- Lymph nodes examined positive
- Relapse Free Status (Months)
- Tumor Stage
- Type of Breast Surgery
- Age at Diagnosis
- Tumor Size



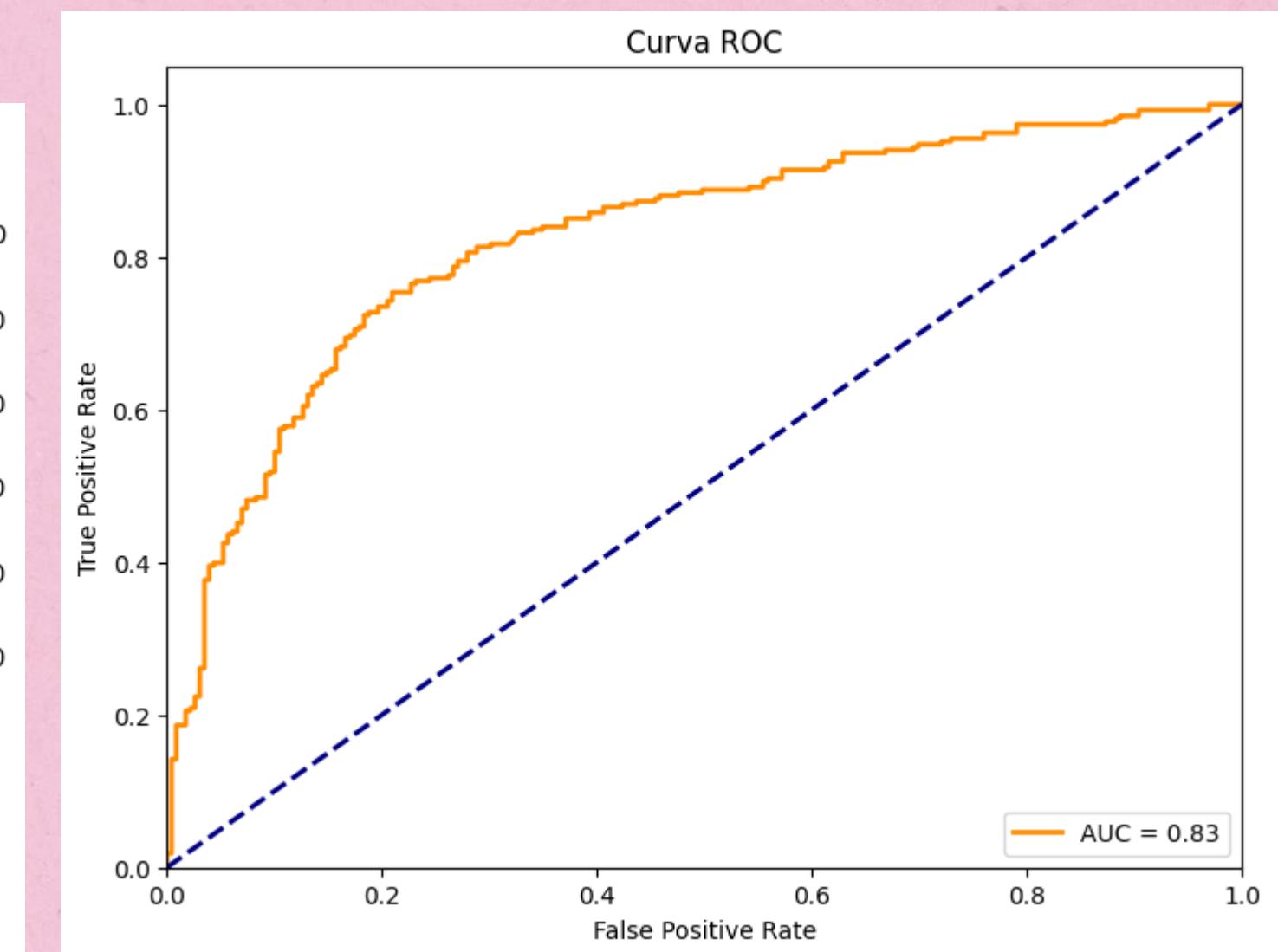
Modelos de Clasificación

LGBM o SVM

Ambas tienen capacidad del modelo para distinguir correctamente entre pacientes con diferentes tiempos de supervivencia

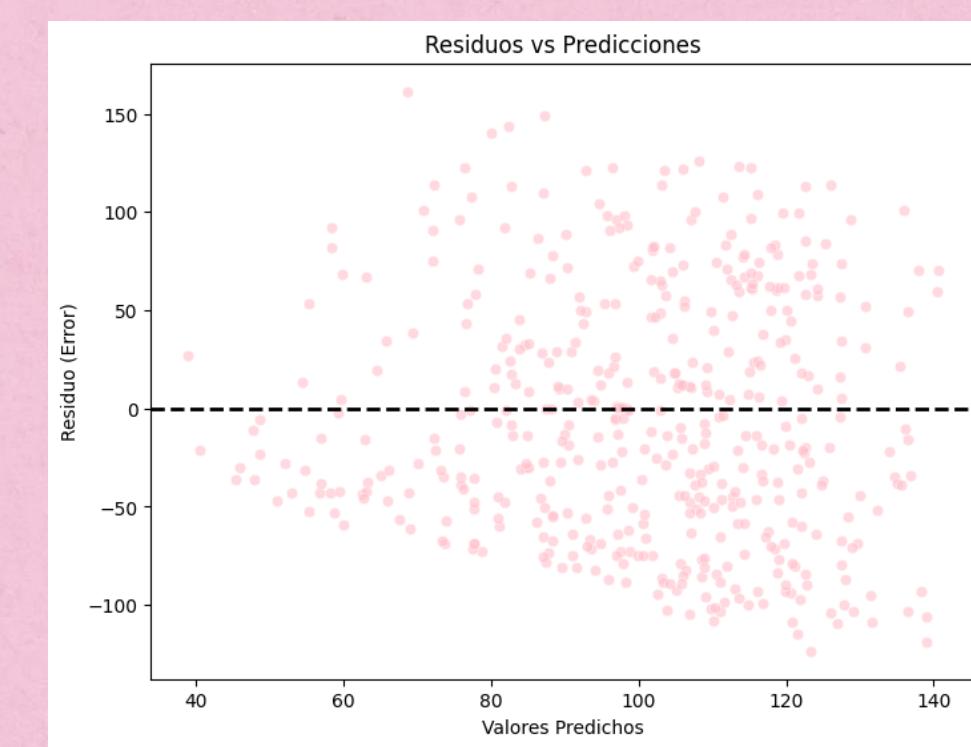
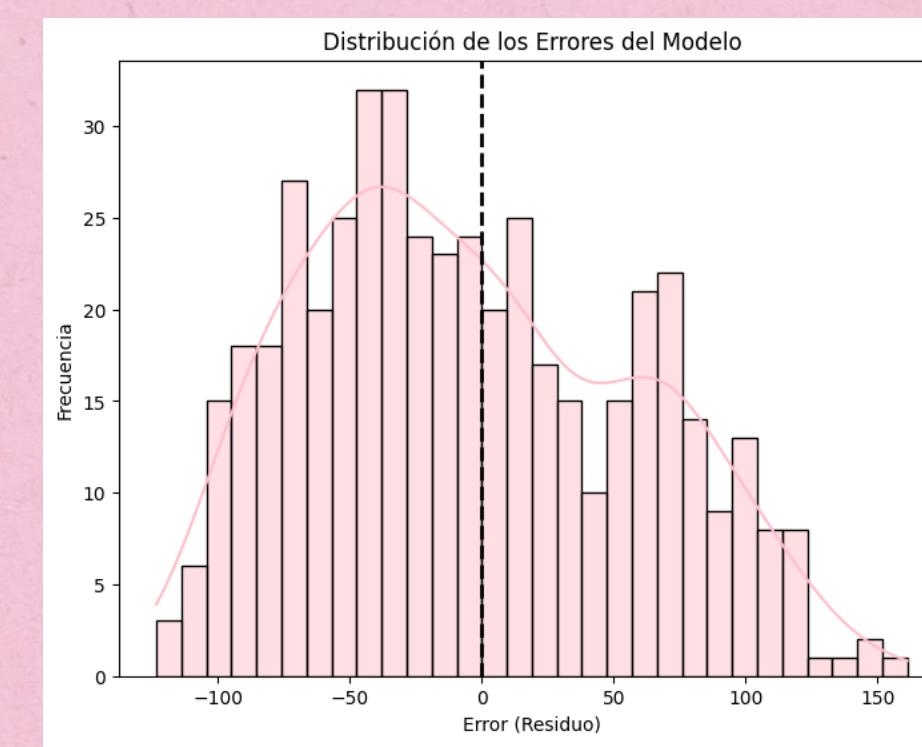
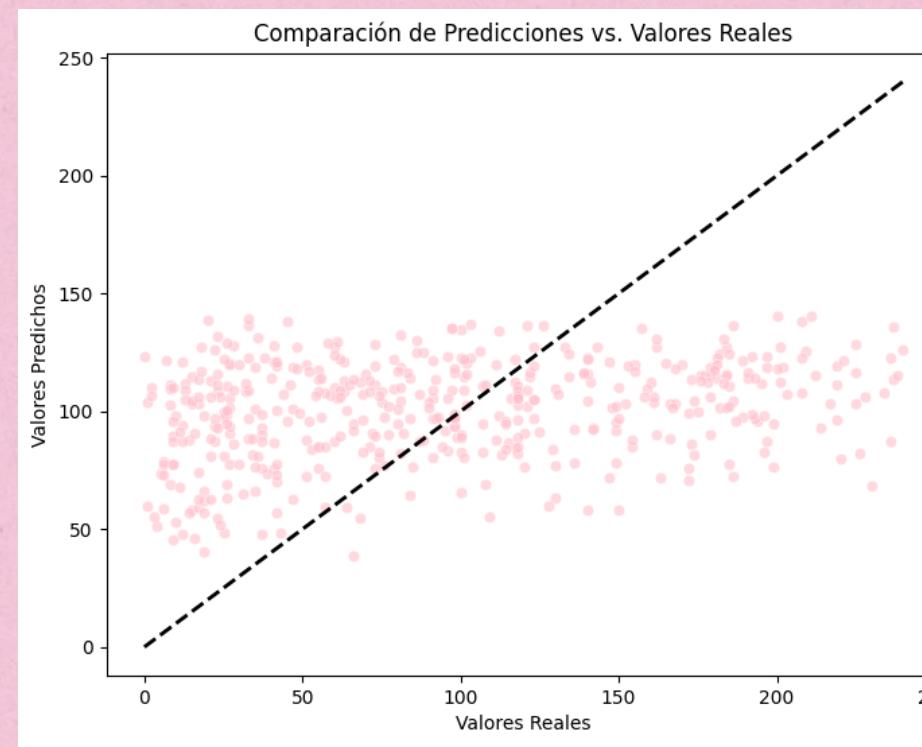
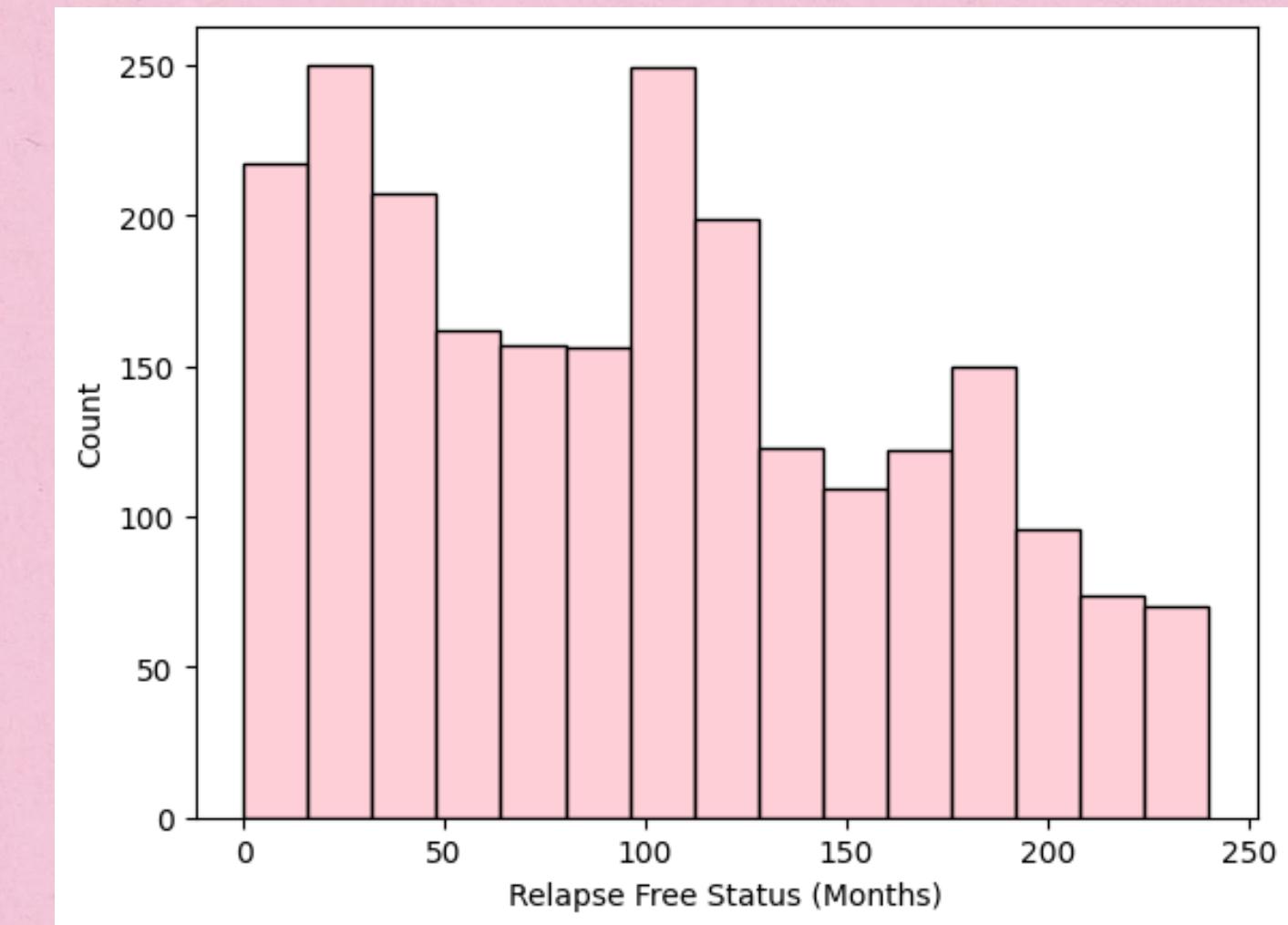


Modelo	Accuracy	F1-Score	ROC-AUC
Random Forest	0.7545	0.7821	0.7492
XGBoost	0.7586	0.7857	0.7532
SVM	0.7525	0.7526	0.8330
Logistic Regression	0.7183	0.7509	0.7124
Gradient Boosting	0.7545	0.7814	0.7495
LightGBM	0.7505	0.7786	0.8331



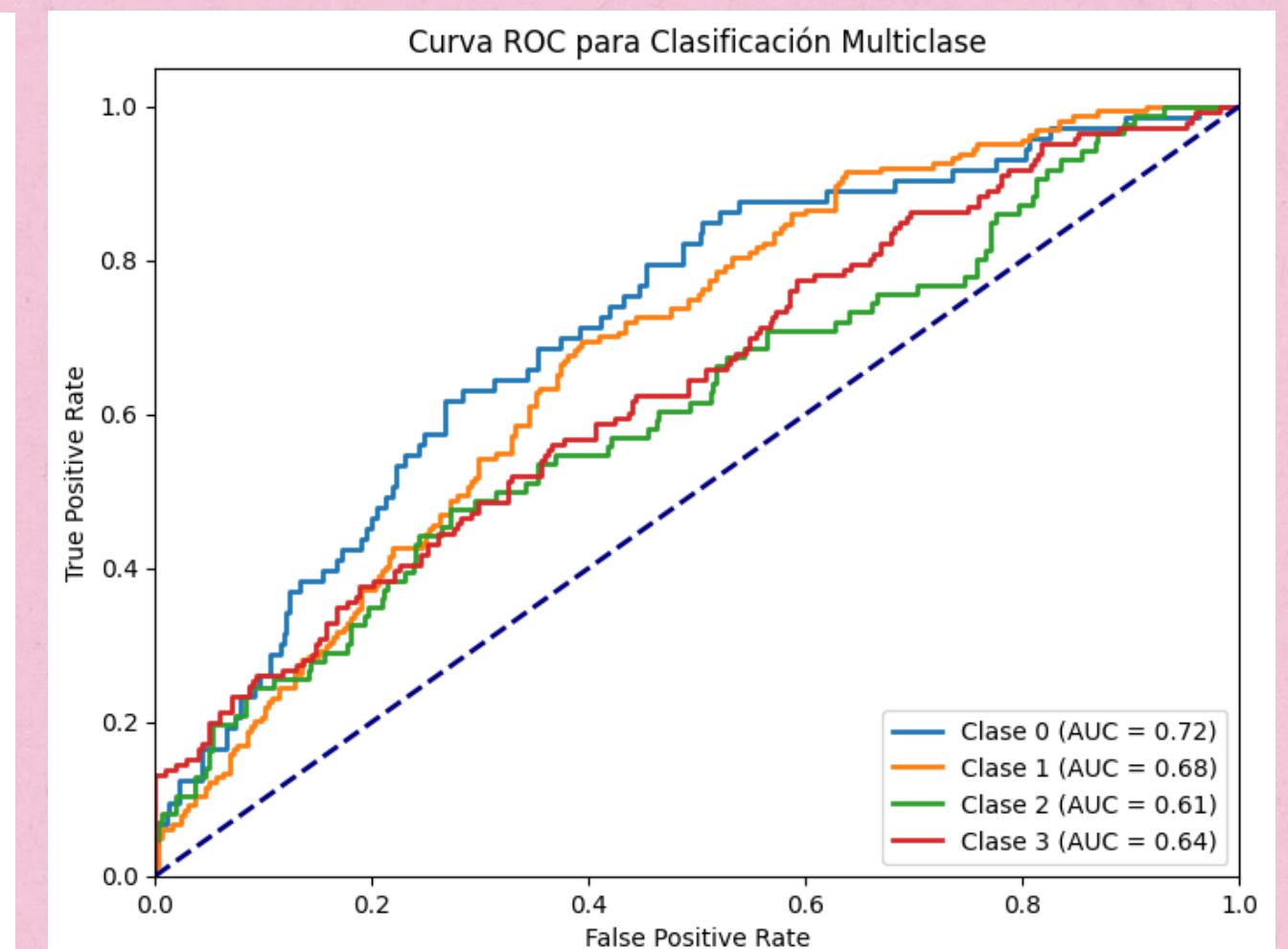
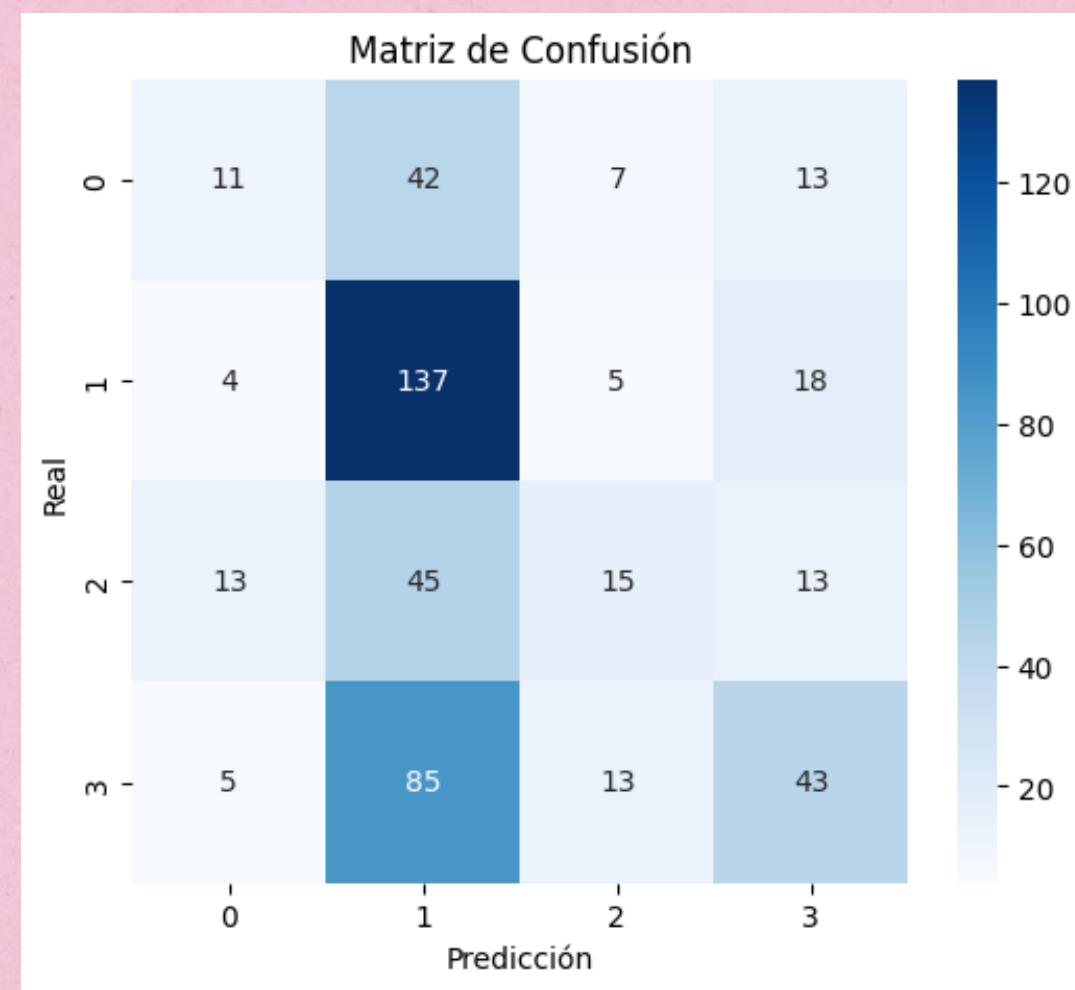
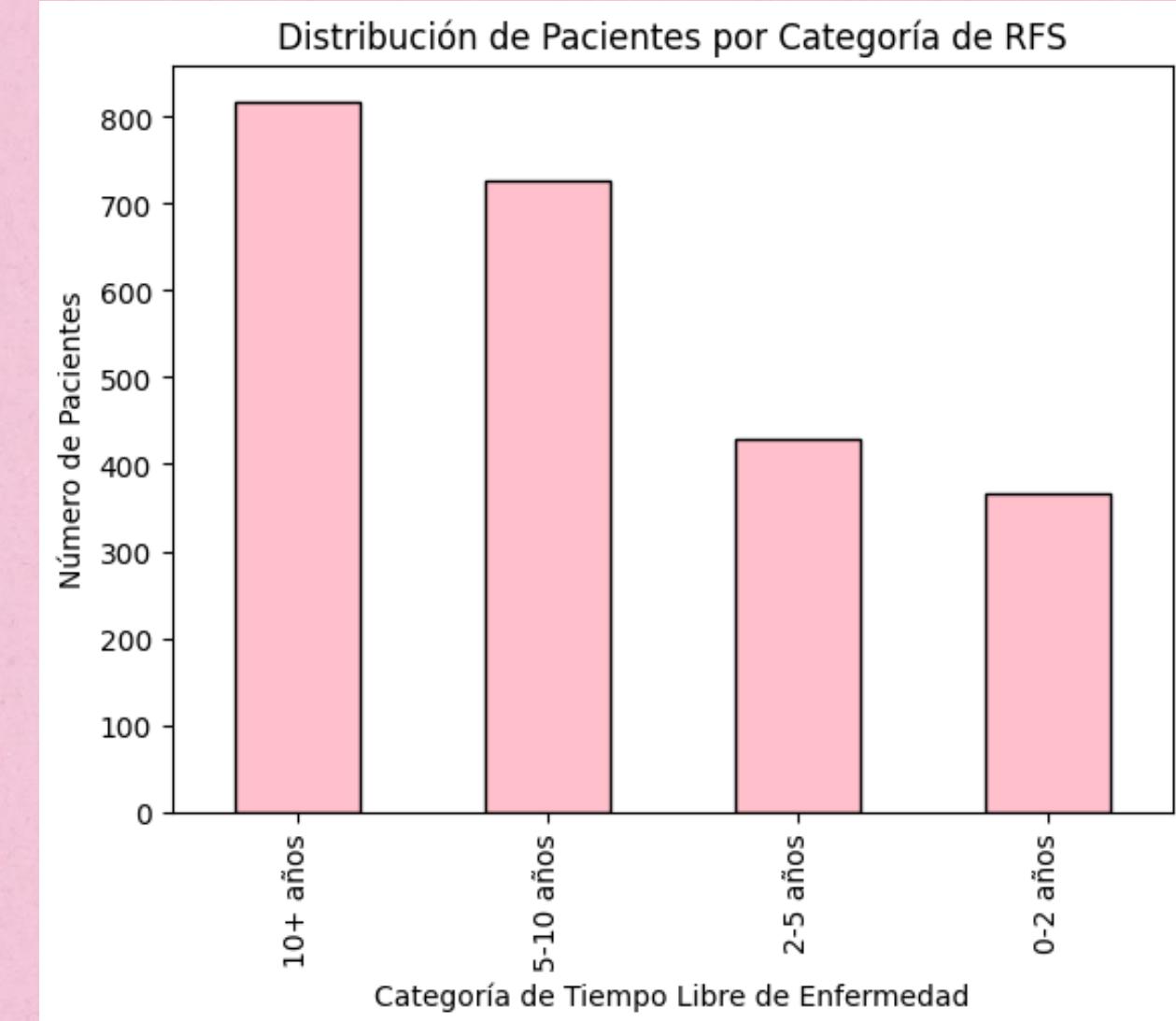
Relapse Free Survival (Months)

Random Forest Regressor



Relapse Free Survival (Months)

Random Forest Classifier



Ventajas de LightGBM

Mejor equilibrio entre rendimiento y eficiencia

Boosting que entrena más rápido y maneja grandes volúmenes de datos.

Alta capacidad de diferenciación (ROC-AUC de 0.8331)

Es el modelo con mejor capacidad para discriminar entre clases sin comprometer precisión o recall.

Buen manejo de datos categóricos y numéricos

Robustez y estabilidad en predicciones

proporcionó predicciones más confiables y menos sesgadas hacia probabilidades intermedias.

Robustez y estabilidad en predicciones

Su capacidad para manejar datos heterogéneos y grandes volúmenes lo hace ideal para ser aplicado en contextos clínicos

Muchas Gracias

Repositorio: https://github.com/Masanadd/ML_breast_cancer.git
