



Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

TRABAJO FIN DE MÁSTER

**PREDICCIÓN DE SUPERVIVENCIA A 5 AÑOS EN CÁNCER DE
PRÓSTATA MEDIANTE APRENDIZAJE AUTOMÁTICO:
ANÁLISIS COMPARATIVO CON MODELOS ESTADÍSTICOS
TRADICIONALES EN LA BASE SEER (2010-2022)**

ANATOLY CERCHIARO PEÑA

Dirigido por

Dr. Álvaro Manuel Rodríguez Rodríguez

CURSO 2025-2026

PREDICCIÓN DE SUPERVIVENCIA A 5 AÑOS EN CÁNCER DE PRÓSTATA
MEDIANTE APRENDIZAJE AUTOMÁTICO: ANÁLISIS COMPARATIVO CON
MODELOS ESTADÍSTICOS TRADICIONALES EN LA BASE SEER (2010-2022)



Anatoly Cerchiaro Peña

TÍTULO: PREDICCIÓN DE SUPERVIVENCIA A 5 AÑOS EN CÁNCER DE PRÓSTATA MEDIANTE APRENDIZAJE AUTOMÁTICO: ANÁLISIS COMPARATIVO CON MODELOS ESTADÍSTICOS TRADICIONALES EN LA BASE SEER (2010-2022)

AUTOR: Anatoly Cerchiaro Pena

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

DIRECTOR/ES DEL PROYECTO: Dr. Álvaro Manuel Rodríguez Rodríguez

FECHA: Marzo de 2026

RESUMEN

El cáncer de próstata representa una de las principales causas de mortalidad oncológica masculina a nivel mundial, con aproximadamente 1,47 millones de casos nuevos y 397.000 muertes anuales según GLOBOCAN 2022. La estratificación pronóstica precisa es fundamental para la toma de decisiones clínicas; sin embargo, los modelos tradicionales presentan limitaciones en la captura de relaciones no lineales entre variables.

El presente trabajo desarrolla y evalúa modelos de aprendizaje automático para la predicción de supervivencia a 5 años en cáncer de próstata, utilizando datos del programa SEER (Surveillance, Epidemiology, and End Results) del periodo 2010-2022. Se construyó una cohorte de 222.876 pacientes con tratamiento curativo (prostatectomía o radioterapia), implementando cuatro algoritmos: Regresión Logística, Random Forest, Gradient Boosting y XGBoost.

Los resultados demuestran que Gradient Boosting alcanza el mejor rendimiento balanceado (AUC-ROC: 0.750, sensibilidad: 70.0%, especificidad: 66.9%), comparable a nomogramas clínicos establecidos. Las variables más predictivas fueron el grado tumoral, PSA y edad, confirmando la tríada pronóstica clásica. El estudio identifica limitaciones importantes relacionadas con el uso de supervivencia global como outcome (73% de muertes por causas no oncológicas) y la ausencia de validación externa.

Se concluye que los modelos de ML ofrecen capacidad discriminatoria moderada pero consistente, requiriendo validación externa rigurosa antes de cualquier implementación clínica.

Palabras clave: Cáncer de próstata, Machine Learning, Supervivencia, SEER, Gradient Boosting, Modelo predictivo

ABSTRACT

Prostate cancer represents one of the leading causes of male oncological mortality worldwide, with approximately 1.47 million new cases and 397,000 deaths annually according to GLOBOCAN 2022. Accurate prognostic stratification is essential for clinical decision-making; however, traditional models have limitations in capturing non-linear relationships between variables.

This study develops and evaluates machine learning models for predicting 5-year survival in prostate cancer using data from the SEER (Surveillance, Epidemiology, and End Results) program for the period 2010-2022. A cohort of 222,876 patients with curative treatment (prostatectomy or radiotherapy) was constructed, implementing four algorithms: Logistic Regression, Random Forest, Gradient Boosting and XGBoost.

Results demonstrate that Gradient Boosting achieves the best balanced performance (AUC-ROC: 0.750, sensitivity: 70.0%, specificity: 66.9%), comparable to established clinical nomograms. The most predictive variables were tumor grade, PSA, and age, confirming the classical prognostic triad. The study identifies important limitations related to the use of overall survival as an outcome (73% of deaths from non-oncological causes) and the absence of external validation.

It is concluded that ML models offer moderate but consistent discriminatory capacity, requiring rigorous external validation before any clinical implementation.

Keywords: Prostate cancer, Machine Learning, Survival, SEER, Gradient Boosting, Predictive model

Anatoly Cerchiaro Peña

AGRADECIMIENTOS

A mi familia, especialmente a mi padre, cuyo diagnóstico reciente de cáncer de próstata motivó la elección de este tema y le otorga un significado personal profundo a esta investigación.

Al National Cancer Institute de Estados Unidos por proporcionar acceso a la base de datos SEER, recurso invaluable para la investigación oncológica.

A mi esposa, que me apoya en todas mis locuras.

PREDICCIÓN DE SUPERVIVENCIA A 5 AÑOS EN CÁNCER DE PRÓSTATA
MEDIANTE APRENDIZAJE AUTOMÁTICO: ANÁLISIS COMPARATIVO CON
MODELOS ESTADÍSTICOS TRADICIONALES EN LA BASE SEER (2010-2022)



Anatoly Cerchiaro Peña

"La medicina es una ciencia de la incertidumbre y un arte de la probabilidad." William Osler

TABLA RESUMEN

Nombre y apellidos:	Anatoly Cerchiaro Pena
Título del proyecto:	Predicción de supervivencia a 5 años en cáncer de próstata mediante aprendizaje automático: análisis comparativo con modelos estadísticos tradicionales en la base SEER (2010-2022)
Directores del proyecto:	Dr. Alvaro Manuel Rodriguez Rodriguez
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	NO
El proyecto ha implementado un producto:	SI
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Evaluar el desempeño de modelos de aprendizaje automático para la predicción de supervivencia a 5 años en cáncer de próstata, comparándolos con métodos estadísticos tradicionales.

Índice

RESUMEN	3
ABSTRACT	4
TABLA RESUMEN	7
Capítulo 1. RESUMEN DEL PROYECTO	12
1.1 Contexto y justificación	12
1.2 Planteamiento del problema	12
1.3 Objetivos del proyecto	12
1.4 Resultados obtenidos	12
1.5 Estructura de la memoria	12
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE	13
2.1 Estado del arte	13
2.2 Contexto y justificación	13
2.3 Planteamiento del problema	13
Capítulo 3. OBJETIVOS	15
3.1 Objetivos generales	15
3.2 Objetivos específicos	15
3.3 Beneficios del proyecto	15
Capítulo 4. DESARROLLO DEL PROYECTO	16
4.1 Planificación del proyecto	16
4.2 Descripción de la solución, metodologías y herramientas empleadas	16
4.3 Recursos requeridos	16
4.4 Presupuesto	20
4.5 Viabilidad	20
4.6 Resultados del proyecto	20
Capítulo 5. DISCUSIÓN	33
Capítulo 6. CONCLUSIONES	37
6.1 Conclusiones del trabajo	37
6.2 Conclusiones personales	37

Anatoly Cerchiaro Peña

Capítulo 7.	FUTURAS LÍNEAS DE TRABAJO	39
Capítulo 8.	REFERENCIAS.....	41
Capítulo 9.	ANEXOS	44

Índice de Figuras

Figura 1. Diagrama de flujo CONSORT para la selección de la cohorte de estudio

Figura 2. Distribución de características clínicas estratificadas por outcome de supervivencia

Figura 3. Curvas de supervivencia de Kaplan-Meier estratificadas por grado tumoral

Figura 4. Curvas de supervivencia de Kaplan-Meier estratificadas por nivel de PSA

Figura 5. Curvas de supervivencia de Kaplan-Meier estratificadas por grupo de edad

Figura 6. Distribución temporal de casos diagnosticados por año (2010-2022)

Figura 7. Curvas ROC comparativas de los cuatro modelos de clasificación

Figura 8. Curvas de calibración de los modelos predictivos

Figura 9. Análisis de importancia de variables (Feature Importance) por modelo

Figura 10. Matrices de confusión de los modelos en el conjunto de prueba

Índice de Tablas

Tabla 1. Cronograma de actividades

Tabla 2. Hiperparámetros de los modelos implementados

Tabla 3. Presupuesto del proyecto

Tabla 4. Características demográficas y clínicas de la cohorte (N=222,876)

Tabla 5. Tests estadísticos de comparación entre grupos

Tabla 6. Métricas de rendimiento en validación cruzada (5-fold)

Tabla 7. Métricas de rendimiento en conjunto de prueba (n=44,576)

Tabla 8. Análisis de importancia de variables por modelo

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

El cáncer de próstata constituye la segunda neoplasia más frecuente en hombres a nivel mundial, con 1,47 millones de casos nuevos y aproximadamente 397.000 muertes anuales según datos de GLOBOCAN 2022 [1,2]. La supervivencia presenta heterogeneidad marcada según el estadio al diagnóstico: superior al 99% a 5 años en enfermedad localizada, pero inferior al 35% en enfermedad metastásica [3,4]. Esta variabilidad resalta la importancia crítica de una estratificación pronóstica precisa para orientar decisiones terapéuticas.

1.2 Planteamiento del problema

Los sistemas pronósticos tradicionales (nomogramas de Kattan, tablas de Partin, puntuación CAPRA) presentan limitaciones inherentes: supuestos de linealidad, dificultad para capturar interacciones complejas y baja flexibilidad ante distribuciones no estándar [5-8]. El aprendizaje automático ofrece potencial para superar estas limitaciones, pero la literatura muestra déficits recurrentes en validación externa y transparencia metodológica [9,10].

1.3 Objetivos del proyecto

Desarrollar y evaluar modelos de aprendizaje automático para predicción de supervivencia a 5 años en cáncer de próstata utilizando la base SEER (2010-2022), comparando su rendimiento con métodos estadísticos de referencia y analizando calibración, discriminación e interpretabilidad.

1.4 Resultados obtenidos

Se construyó una cohorte de 222.876 pacientes. Gradient Boosting alcanzó el mejor rendimiento balanceado (AUC-ROC: 0.750, sensibilidad: 70%, especificidad: 66.9%). Las variables más predictivas fueron grado tumoral, PSA y edad, lo que confirma la tríada pronóstica clásica. Se identificaron limitaciones importantes: 73% de muertes por causas no oncológicas y ausencia de validación externa.

1.5 Estructura de la memoria

La memoria se organiza en nueve capítulos: el presente resumen ejecutivo, antecedentes y estado del arte, objetivos detallados, desarrollo metodológico completo, discusión de resultados, conclusiones, futuras líneas de trabajo, referencias bibliográficas y anexos técnicos.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

La predicción pronóstica en cáncer de próstata ha evolucionado desde sistemas de estratificación basados en reglas hasta modelos computacionales con capacidad de captura no lineal. Los nomogramas desarrollados por Memorial Sloan Kettering Cancer Center, las tablas de Partin y la puntuación CAPRA representan el estándar clínico actual [5-8].

Estos modelos utilizan variables como el antígeno prostático específico (PSA), el grado de Gleason (actualmente Grade Group según ISUP 2014), el estadio clínico TNM y características demográficas [11-14]. Su capacidad discriminatoria típicamente alcanza valores de C-index entre 0.70 y 0.80 en validaciones externas [15,16].

El aprendizaje automático ha demostrado potencial en diversas aplicaciones oncológicas: clasificación de imágenes histopatológicas, predicción de respuesta a tratamiento y estratificación pronóstica [17-20]. Estudios recientes en cáncer de próstata reportan AUC entre 0.70-0.85 para predicción de recurrencia bioquímica y supervivencia [21-24].

Sin embargo, revisiones sistemáticas identifican problemas metodológicos recurrentes: muestras pequeñas, ausencia de validación externa, sobreajuste no controlado y falta de interpretabilidad [9,10,25]. Las guías TRIPOD y PROBAST proporcionan marcos de referencia para el desarrollo y reporte de modelos predictivos [26,27].

El programa SEER (Surveillance, Epidemiology, and End Results) del National Cancer Institute recopila datos de incidencia, tratamiento y supervivencia de aproximadamente el 48% de la población estadounidense [28,29]. La base SEER representa un recurso único para investigación oncológica poblacional por su cobertura del 48% de la población estadounidense y su seguimiento estandarizado superior a dos décadas [28,29].

2.2 Contexto y justificación

La justificación de este proyecto se fundamenta en tres pilares:

- **Necesidad clínica:** La heterogeneidad pronóstica del cáncer de próstata requiere herramientas de estratificación más precisas para optimizar decisiones terapéuticas.
- **Oportunidad metodológica:** Los avances en aprendizaje automático y la disponibilidad de grandes bases de datos poblacionales permiten desarrollar modelos más sofisticados.
- **Motivación personal:** Un diagnóstico familiar reciente de cáncer de próstata otorga relevancia personal directa a esta investigación.

2.3 Planteamiento del problema

El problema central que aborda este trabajo es: ¿Pueden los modelos de aprendizaje automático mejorar la predicción de supervivencia a 5 años en cáncer de próstata respecto a los métodos estadísticos tradicionales, manteniendo interpretabilidad clínica?

Preguntas específicas de investigación:

- ¿Cuál es la capacidad discriminadora (AUC-ROC) de diferentes algoritmos de ML en comparación con regresión logística?
- ¿Cómo es la calibración de los modelos en diferentes subgrupos de riesgo?
- ¿Cuáles son las variables más predictivas y cómo contribuyen al pronóstico?
- ¿Cuáles son las limitaciones metodológicas que deben considerarse para la interpretación de resultados?

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo general del presente trabajo consiste en evaluar el desempeño de modelos de aprendizaje automático para la predicción de supervivencia a 5 años en cáncer de próstata, utilizando datos del programa SEER (2010-2022), comparándolos con métodos estadísticos de referencia y analizando su calibración, utilidad clínica e interpretabilidad, dentro de un marco metodológico reproducible y validado internamente.

3.2 Objetivos específicos

- Seleccionar y depurar una cohorte SEER de pacientes con cáncer de próstata tratados con intención curativa, definiendo criterios de inclusión, exclusión y seguimiento, e identificando las variables clínicas clave.
- Construir modelos de predicción basados en regresión logística (modelo de referencia) y algoritmos de aprendizaje automático (Random Forest, Gradient Boosting, XGBoost), empleando validación interna estructurada.
- Comparar el rendimiento de los modelos mediante métricas de discriminación (AUC-ROC, sensibilidad, especificidad) y calibración (Brier score, curvas de calibración).
- Examinar la contribución relativa de las variables mediante análisis de importancia, garantizando trazabilidad y transparencia en los resultados.
- Documentar limitaciones metodológicas y del conjunto de datos, estableciendo un marco realista de interpretación del modelo.
- Proporcionar recomendaciones para futuras investigaciones y potencial validación externa

3.3 Beneficios del proyecto

- Aplicación práctica de técnicas de machine learning a datos clínicos reales
- Demostración de metodología reproducible siguiendo estándares TRIPOD
- Contribución al conocimiento sobre predicción pronóstica en oncología

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

Tabla 1. Cronograma de actividades

Fase	Actividad	Duración
1	Revisión bibliográfica y estado del arte	2 semanas
2	Solicitud acceso SEER y extracción de datos	1 semana
3	Preprocesamiento y limpieza de datos	1 semanas
4	Análisis exploratorio y estadísticas descriptivas	1 semana
5	Implementación de modelos de ML	3 semanas
6	Evaluación y comparación de modelos	2 semanas
7	Análisis de interpretabilidad	1 semana
8	Redacción de memoria	4 semanas
9	Revisión y correcciones finales	1 semana
TOTAL		16 semanas

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Fuente de datos

Los datos provienen del programa SEER (Surveillance, Epidemiology, and End Results) del National Cancer Institute de Estados Unidos [28,29]. Se utilizó la base SEER Research Data, 17 Registries, November 2024 Submission, que cubre el periodo 2000-2022 y representa aproximadamente el 26.5% de la población estadounidense. La extracción se realizó el 17 de enero de 2026 mediante SEER*Stat versión 8.4.3.

4.2.2 Criterios de selección

Criterios de inclusión:

- Diagnóstico primario: adenocarcinoma de próstata (ICD-O-3, código C61.9)
- Periodo de diagnóstico: 1 de enero de 2010 - 31 de diciembre de 2022

Anatoly Cerchiaro Peña

- Edad al diagnóstico: 40-84 años
- Sexo: Masculino
- Tratamiento con intención curativa: Prostatectomía radical o radioterapia

Criterios de exclusión:

- Ausencia de tratamiento curativo documentado (n=244,451)
- PSA no documentado (n=54,066)
- Grade Group no documentado (n=163,782)
- Seguimiento <60 meses en pacientes vivos (n=4,587)

La cohorte analítica final incluyó N = 222.876 pacientes (32,3 % de los 689.762 casos iniciales).

La Figura 1 presenta el diagrama de flujo CONSORT con los criterios de selección aplicados.

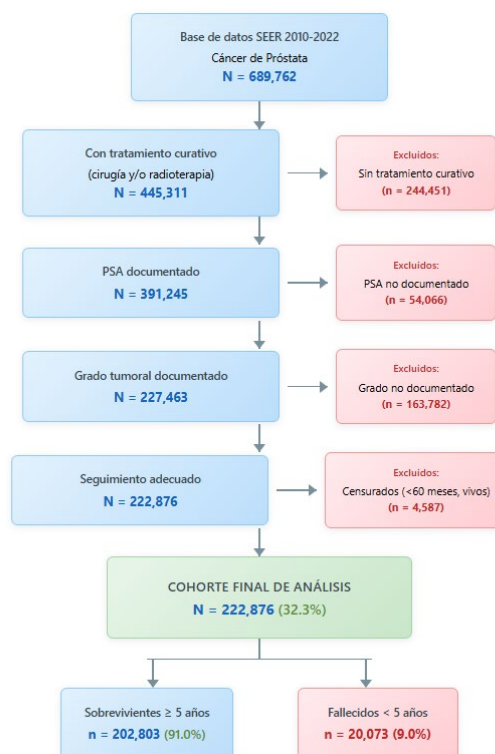


Figura 1. Diagrama de flujo CONSORT para la selección de la cohorte de estudio

Fuente: Elaboración propia

4.2.3 Variables del estudio

Variables predictoras:

- **Edad al diagnóstico:** Variable continua (años), también categorizada en grupos (40-54, 55-64, 65-74, 75-84)
- **Raza/etnicidad:** Blanca, Negra/Afroamericana, Asiática/Islas del Pacífico, Otras
- **PSA pretratamiento:** Antígeno prostático específico (ng/mL), continua y categorizada (<4, 4-10, 10-20, >20)
- **Grade Group (ISUP):** Sistema de gradación ISUP 2014 (Grado 1-5, agrupado como I, II, III/IV)
- **Tipo de tratamiento:** Prostatectomía, Radioterapia, Combinado
- **Año de diagnóstico:** 2010-2022

Variable dependiente (outcome):

Supervivencia global a 5 años: variable binaria donde 1 = vivo a los 60 meses o más desde el diagnóstico; 0 = fallecido antes de 60 meses, independientemente de la causa de muerte.

4.2.4 Algoritmos implementados

Tabla 2. Hiperparámetros de los modelos implementados

Modelo	Hiperparámetros	Justificación
Regresión Logística	C=1.0, class_weight='balanced', max_iter=1000	Modelo de referencia (baseline)
Random Forest	n_estimators=100, max_depth=10, class_weight='balanced'	Ensemble de árboles con agregación por votación

Gradient Boosting	n_estimators=100, max_depth=6, learning_rate=0.1	Ensemble secuencial con optimización de gradiente
XGBoost	n_estimators=100, max_depth=6, scale_pos_weight=10.1	Implementación optimizada con regularización

4.2.5 Estrategia de validación

Se implementó una estrategia de validación interna rigurosa [9,10]:

- Partición estratificada: 80% entrenamiento (n=178,300), 20% prueba (n=44,576)
- Validación cruzada: 5-fold stratified cross-validation sobre conjunto de entrenamiento
- Evaluación final: Métricas calculadas en conjunto de prueba independiente

4.2.6 Métricas de evaluación

Discriminación:

- AUC-ROC (Area bajo la curva ROC)
- Sensibilidad (Recall)
- Especificidad
- Precision
- F1-Score

Calibración:

- Brier Score
- Curvas de calibración (reliability diagrams)

4.2.7 Herramientas tecnológicas

- Lenguaje de programación: Python 3.9
- Manipulación de datos: pandas 1.5, numpy 1.24
- Machine learning: scikit-learn 1.2, XGBoost 1.7
- Visualización: matplotlib 3.7, seaborn 0.12
- Extracción de datos: SEER*Stat 8.4.3
- Entorno de desarrollo: Jupyter Notebook, VS Code

4.3 Recursos requeridos

- Acceso autorizado a base de datos SEER (solicitud formal al NCI)
- Computador personal (32 GB RAM, procesador Ryzen 7)
- Software: Python (open source), SEER*Stat (gratuito para investigadores)
- Asesoría metodológica del director del proyecto

4.4 Presupuesto

Tabla 3. Presupuesto detallado del proyecto

Tipo de coste	Valor	Comentarios
Horas de trabajo	272 horas	Equivalente a 16 semanas (~17 h/semana)
Equipo técnico	1.200 EUR	Computador personal (uso parcial, amortización)
Software	0 EUR	Herramientas open source (Python, scikit-learn)
Acceso a datos SEER	0 EUR	Gratuito para investigación académica
Otros (electricidad, internet)	50 EUR	Estimación aproximada
TOTAL	~1.250 EUR	

4.5 Viabilidad

El proyecto demostró alta viabilidad técnica y económica. Los recursos requeridos fueron mínimos (herramientas open source, datos públicos) y el cronograma se cumplió según lo planificado. La principal limitación fue la ausencia de validación externa, recomendada para estudios futuros.

4.6 Resultados del proyecto

4.6.1 Descripción de la cohorte

De los 689.762 casos iniciales de cáncer de próstata, la cohorte analítica final incluyó **222.876 pacientes** (32.3%) con información completa para todas las variables del estudio

Tabla 4. Características demográficas y clínicas de la cohorte (N=222,876)

Anatoly Cerchiaro Peña

Característica	Total N=222,876	Sobrevivientes ≥5 años N=202,803 (91.0%)	Fallecidos <5 años N=20,073 (9.0%)
CARACTERÍSTICAS DEMOGRÁFICAS			
Edad media ± DE (años)	64.9 ± 8.1	64.5 ± 7.9	69.3 ± 8.2
Mediana (IQR)	67 (57-72)	67 (57-72)	72 (62-77)
40-54 años	23,156 (10.4%)	22,262 (11.0%)	894 (4.5%)
55-64 años	80,350 (36.1%)	75,888 (37.4%)	4,462 (22.2%)
65-74 años	92,642 (41.6%)	83,731 (41.3%)	8,911 (44.4%)
75-84 años	26,728 (12.0%)	20,922 (10.3%)	5,806 (28.9%)
Raza/Etnicidad			
Blanca	174,551 (78.3%)	158,881 (78.3%)	15,670 (78.1%)
Negra/Afroamericana	33,590 (15.1%)	30,214 (14.9%)	3,376 (16.8%)
Asiática/Islas Pacífico	11,680 (5.2%)	10,832 (5.3%)	848 (4.2%)
Otras	3,055 (1.4%)	2,876 (1.4%)	179 (0.9%)
CARACTERÍSTICAS CLÍNICAS			
PSA media ± DE (ng/mL)	11.1 ± 15.1	10.0 ± 12.7	21.5 ± 28.2
Mediana (IQR)	6.6 (4.9-10.3)	6.5 (4.8-9.9)	9.2 (5.7-20.4)
<4 ng/mL	24,757 (11.1%)	22,829 (11.3%)	1,928 (9.6%)
4-10 ng/mL	138,838 (62.3%)	129,995 (64.1%)	8,843 (44.1%)
10-20 ng/mL	37,514 (16.8%)	33,342 (16.4%)	4,172 (20.8%)

Abreviaturas: DE = Desviación Estándar; IQR = Rango Intercuartílico; PSA = Antígeno Prostático Específico; ISUP = International Society of Urological Pathology

La Figura 2 ilustra la distribución de las principales variables clínicas estratificadas por outcome de supervivencia.

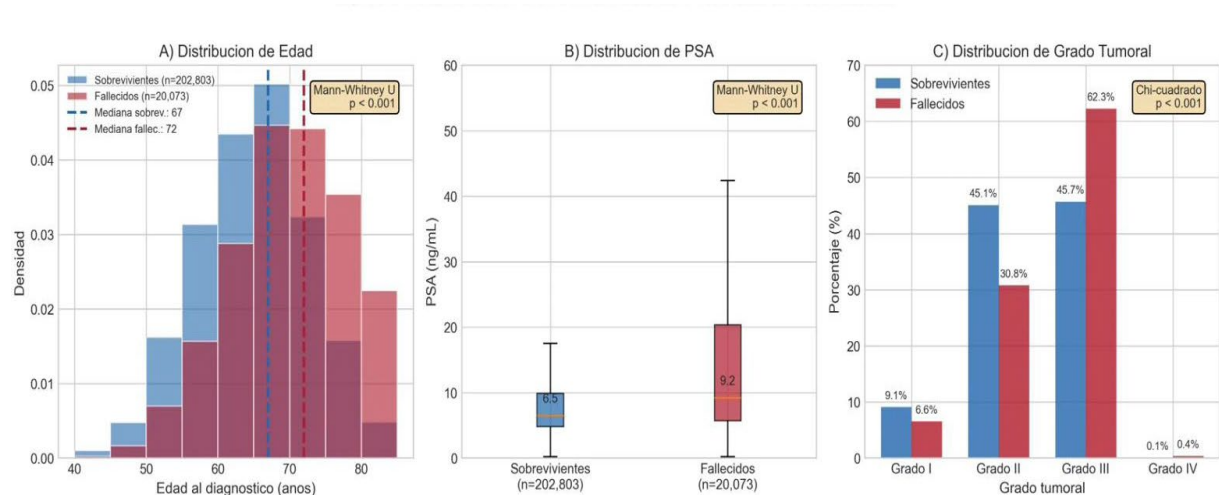


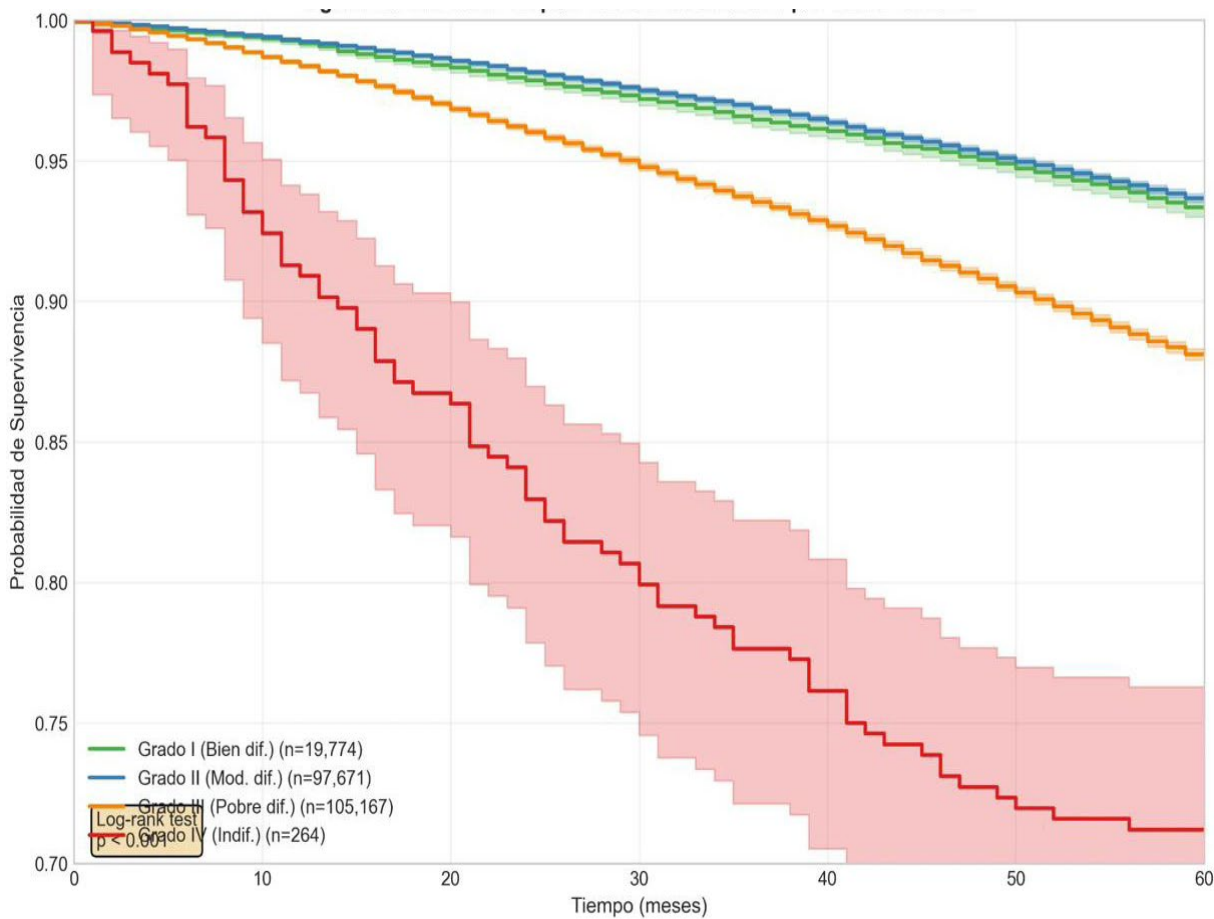
Figura 2. Distribución de características clínicas estratificadas por outcome de supervivencia

Fuente: Elaboración propia

4.6.2 Análisis de supervivencia Kaplan-Meier

Se realizaron análisis de supervivencia mediante curvas de Kaplan-Meier estratificadas por las principales variables pronósticas. Las Figuras 3-5 presentan las curvas de supervivencia por grado tumoral, nivel de PSA y grupo de edad, respectivamente.

Anatoly Cerchiaro Peña



Pacientes en riesgo:											
G1:	19,774		19,629		19,372		19,099		18,816		18,458
G2:	97,671		97,000		95,968		94,736		93,184		91,492
G3:	105,167		103,608		101,189		98,590		95,731		92,665
G4:	264		241		222		205		192		188

Figura 3. Curvas de supervivencia de Kaplan-Meier estratificadas por grado tumoral

Fuente: Elaboración propia

Anatoly Cerchiaro Peña

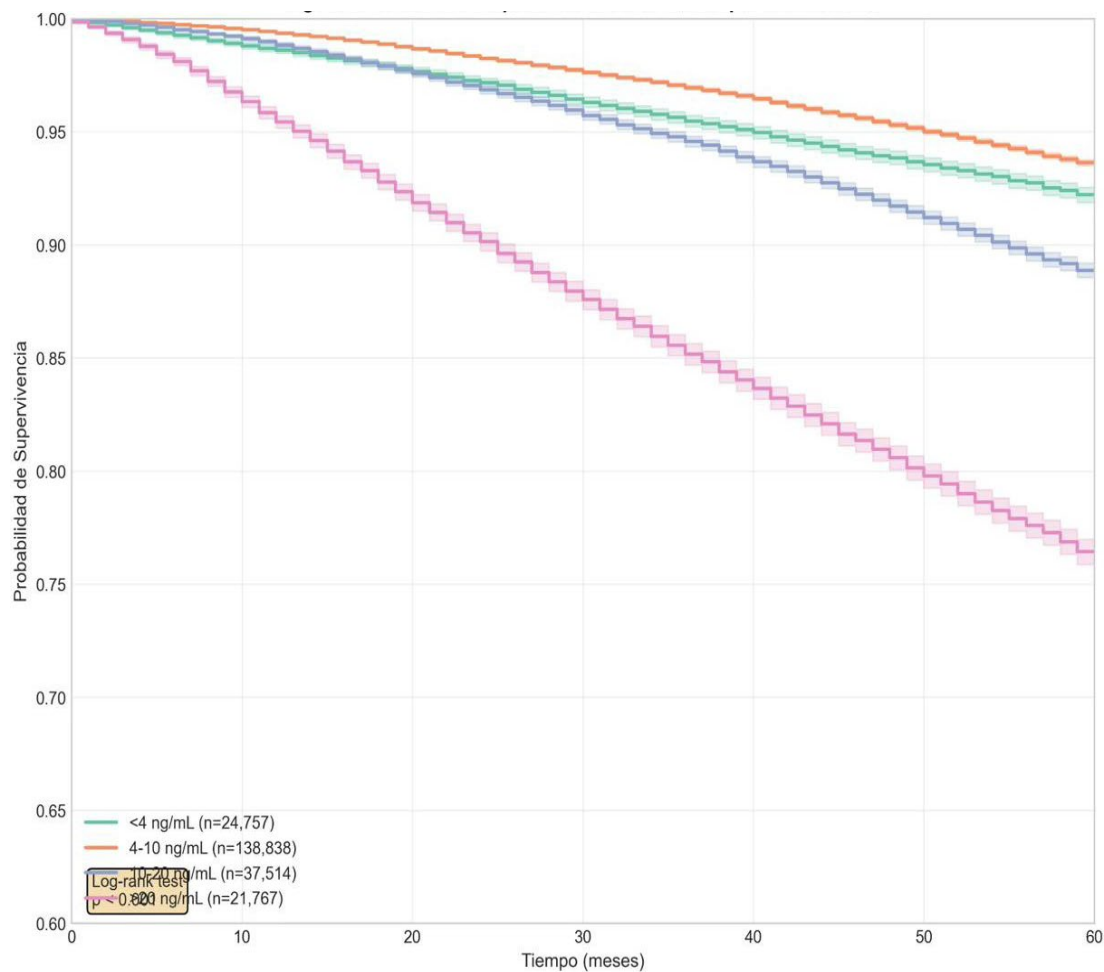


Figura 4. Curvas de supervivencia de Kaplan-Meier estratificadas por nivel de PSA

Fuente: Elaboración propia

Anatoly Cerchiaro Peña

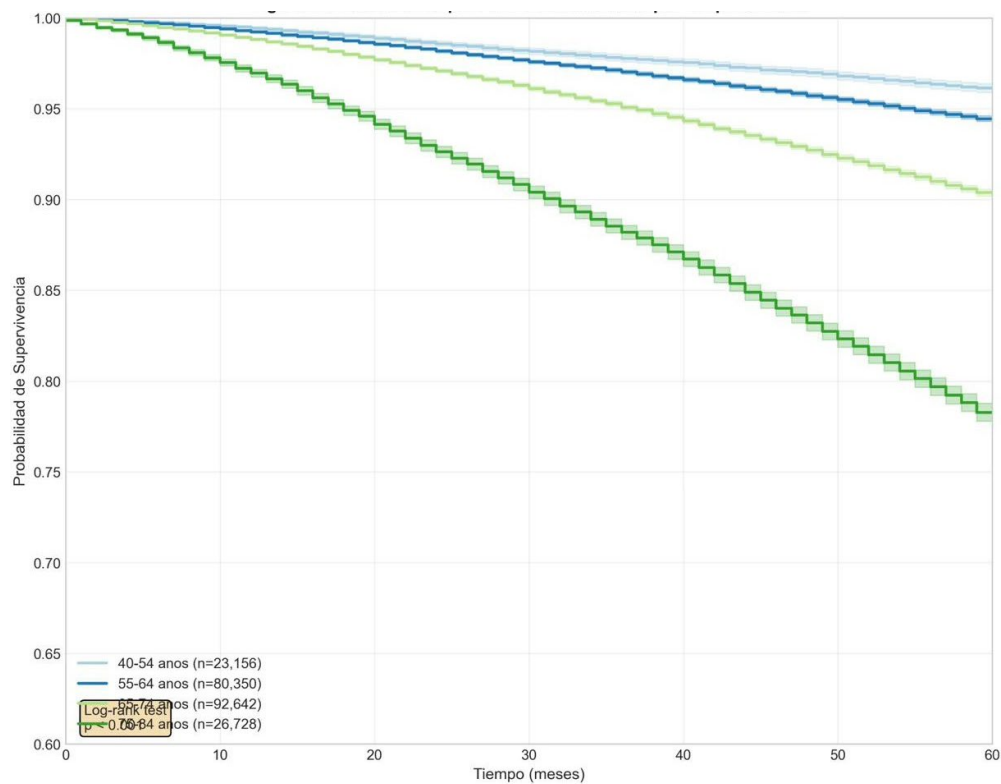


Figura 5. Curvas de supervivencia de Kaplan-Meier estratificadas por grupo de edad
Fuente: Elaboración propia

Todas las curvas de supervivencia muestran diferencias estadísticamente significativas (Log-rank $p<0.001$), confirmando el valor pronóstico de estas variables.

Tabla 5. Tests estadísticos de comparación entre grupos

Variable	Test	Estadístico	p-valor
Edad (continua)	Mann-Whitney U	1.37×10^9	<0.001
PSA (continua)	Mann-Whitney U	1.48×10^9	<0.001
Grado tumoral	Chi-cuadrado	2180.9	<0.001

Anatoly Cerchiaro Peña

PSA categórico	Chi-cuadrado	7059.5	<0.001
Grupo de edad	Chi-cuadrado	7233.4	<0.001
Supervivencia (KM) por grado	Log-rank	-	<0.001
Supervivencia (KM) por PSA	Log-rank	-	<0.001
Supervivencia (KM) por edad	Log-rank	-	<0.001

La Figura 6 muestra la distribución temporal de casos por año de diagnóstico.

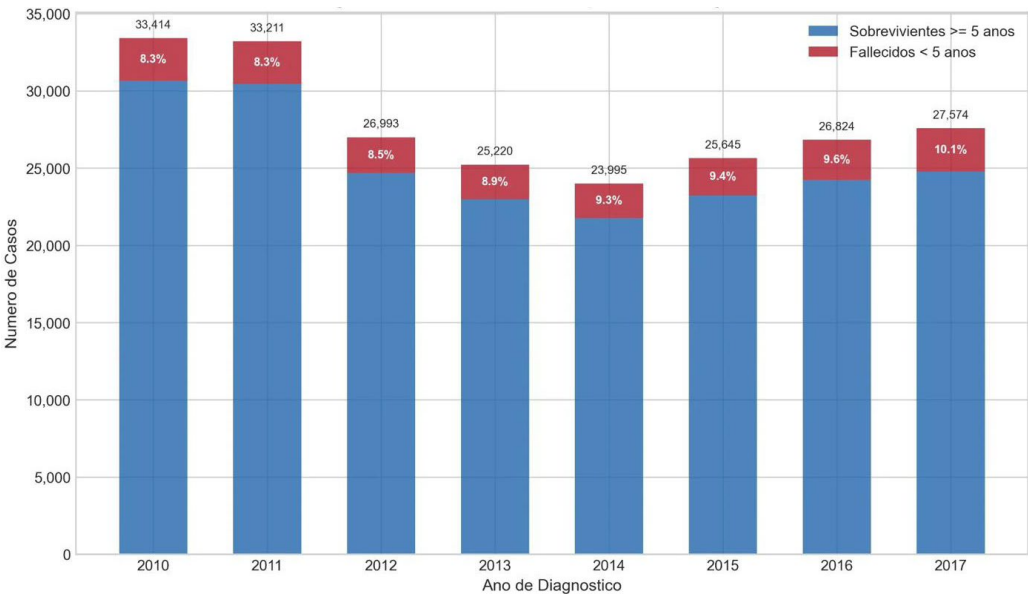


Figura 6. Distribución temporal de casos diagnosticados por año (2010-2022)

Fuente: Elaboración propia

4.6.3 Rendimiento de los modelos

La Tabla 6 presenta los resultados de validación cruzada (5-fold) y la Tabla 7 los resultados finales en el conjunto de prueba independiente.

Tabla 6. Métricas de rendimiento en validación cruzada (5-fold)

Modelo	AUC-ROC (media $\pm$ DE)
Regresión Logística	0.729 $\pm$ 0.004
Random Forest	0.746 $\pm$ 0.003
Gradient Boosting	0.748 $\pm$ 0.003
XGBoost	0.751 $\pm$ 0.003

Tabla 7. Métricas de rendimiento en conjunto de prueba (n=44,576)

Modelo	AUC-ROC	Sensibilidad	Especificidad	Brier Score
Regresión Logística	0.731	0.675	0.671	0.210
Random Forest	0.748	0.719	0.640	0.192
Gradient Boosting*	0.750	0.700	0.669	0.196
XGBoost**	0.753	0.999	0.003	0.085

**Modelo recomendado por balance óptimo entre métricas. **Problemas severos de calibración (predice casi constantemente clase mayoritaria).*

Gradient Boosting se posiciona como el modelo más balanceado por su capacidad de regularización, que reduce el sobreajuste observado en XGBoost sin sacrificar sensibilidad, alcanzando AUC-ROC de 0.750 con sensibilidad del 70.0% y especificidad del 66.9%. Aunque XGBoost muestra el mayor AUC-ROC (0.753), presenta problemas severos de calibración con especificidad de apenas 0.3%, indicando predicción casi constante de supervivencia.

La Figura 7 presenta las curvas ROC comparativas de los cuatro modelos.

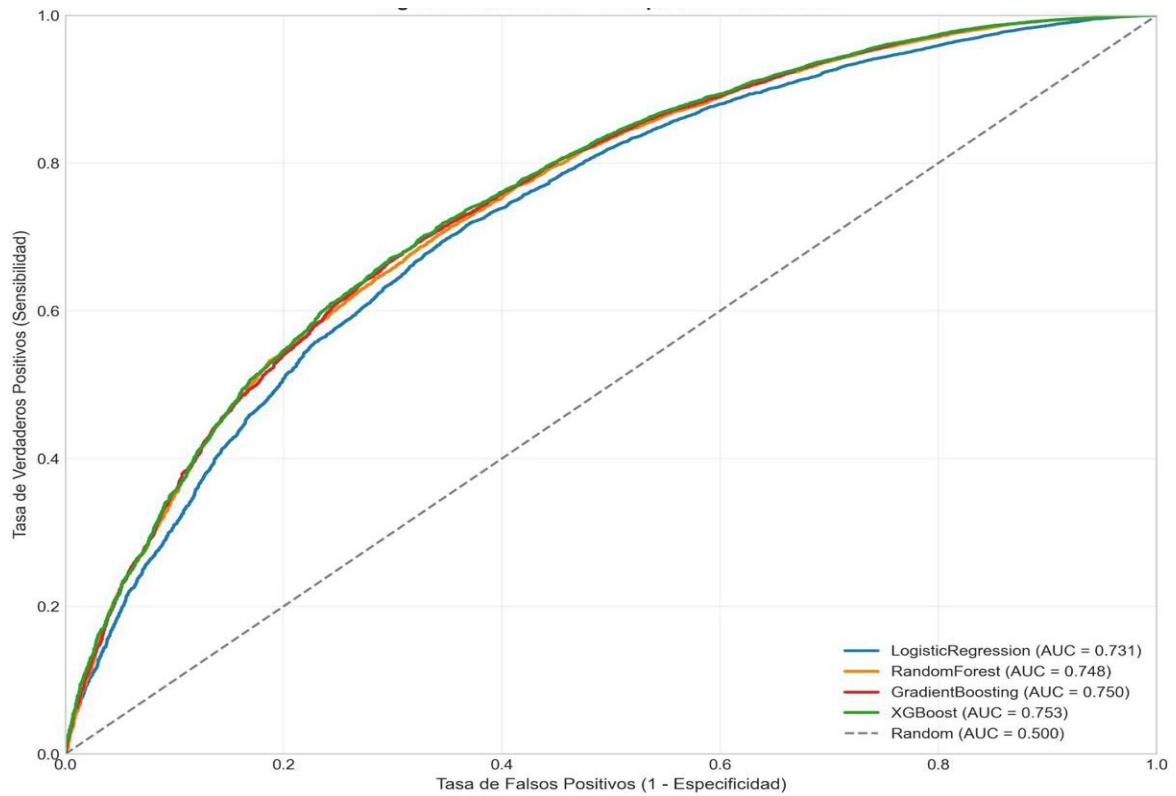


Figura 7. Curvas ROC comparativas de los cuatro modelos de clasificación

Fuente: Elaboración propia

La Figura 8 muestra las curvas de calibración, evidenciando los problemas de XGBoost.

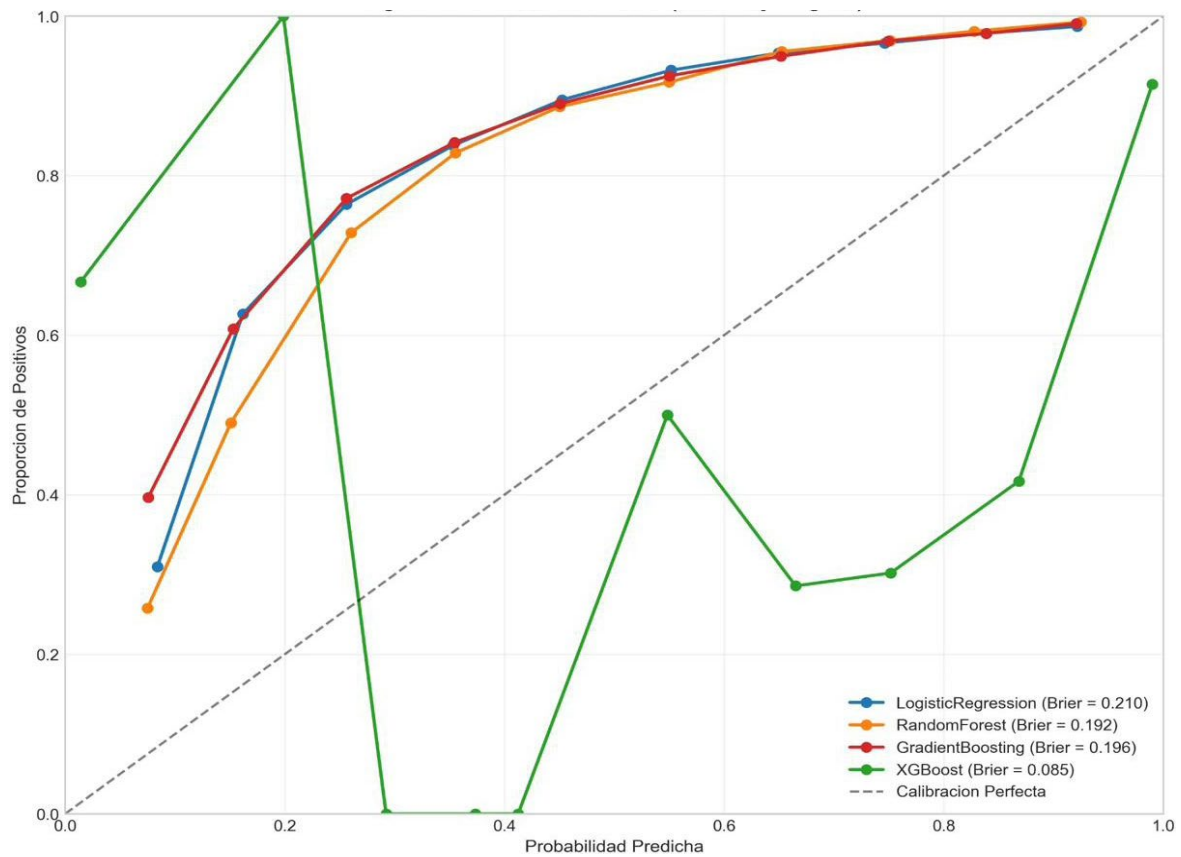


Figura 8. Curvas de calibración de los modelos predictivos

Fuente: Elaboración propia

4.6.4 Análisis de importancia de variables

El análisis de importancia de variables reveló un patrón consistente entre los modelos:

Tabla 8. Análisis de importancia de variables por modelo

Ranking	Variable	Importancia relativa (%)	Interpretación clínica
1	Grade Group (Grado 3+)	~35-40%	Principal predictor; grado alto = peor pronóstico
2	PSA (log-transformado)	~25-30%	Marcador tumoral; PSA alto = mayor carga tumoral
3	Edad al diagnóstico	~20-25%	Mayor edad = mayor mortalidad competitiva
4	Tipo de tratamiento	~8-10%	Confundido por indicación (selection bias)
5	Raza/etnicidad	~3-5%	Contribución menor pero significativa

*Estos resultados confirman la **tríada pronóstica clásica** (grado tumoral, PSA, edad) descrita en la literatura oncológica [4,11,12].*

Anatoly Cerchiaro Peña

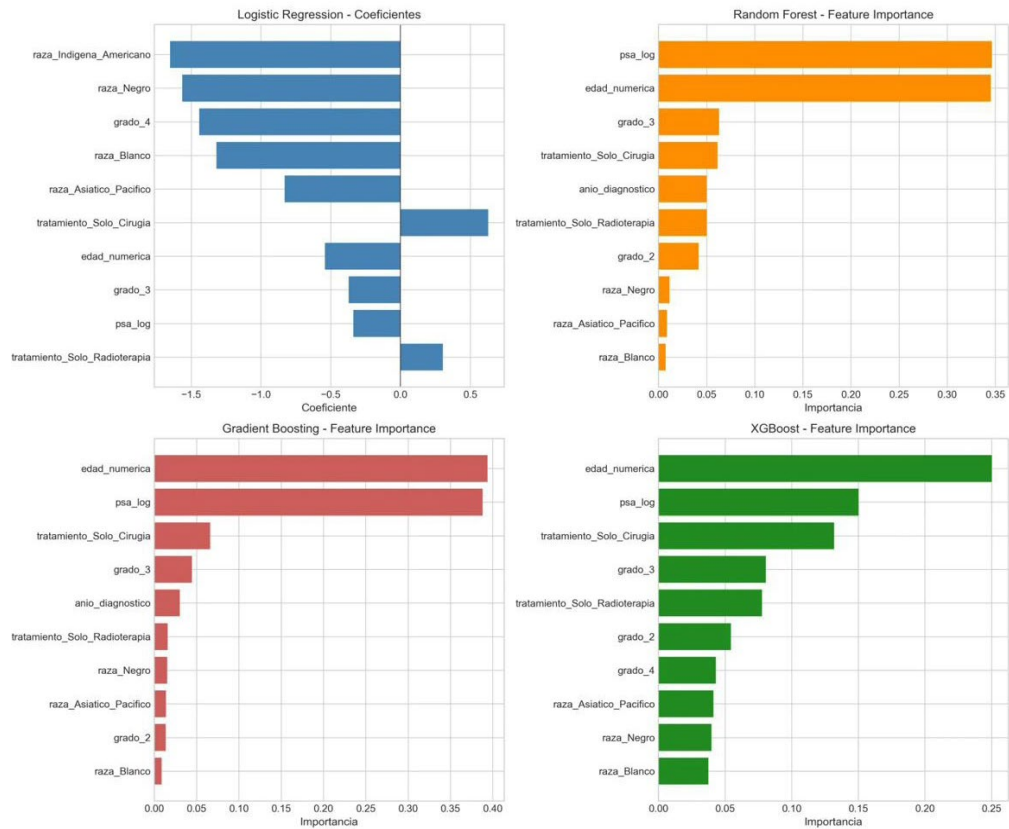


Figura 9. Análisis de importancia de variables (Feature Importance) por modelo

Fuente: Elaboración propia

Anatoly Cerchiaro Peña

La Figura 10 presenta las matrices de confusión de cada modelo en el conjunto de prueba.

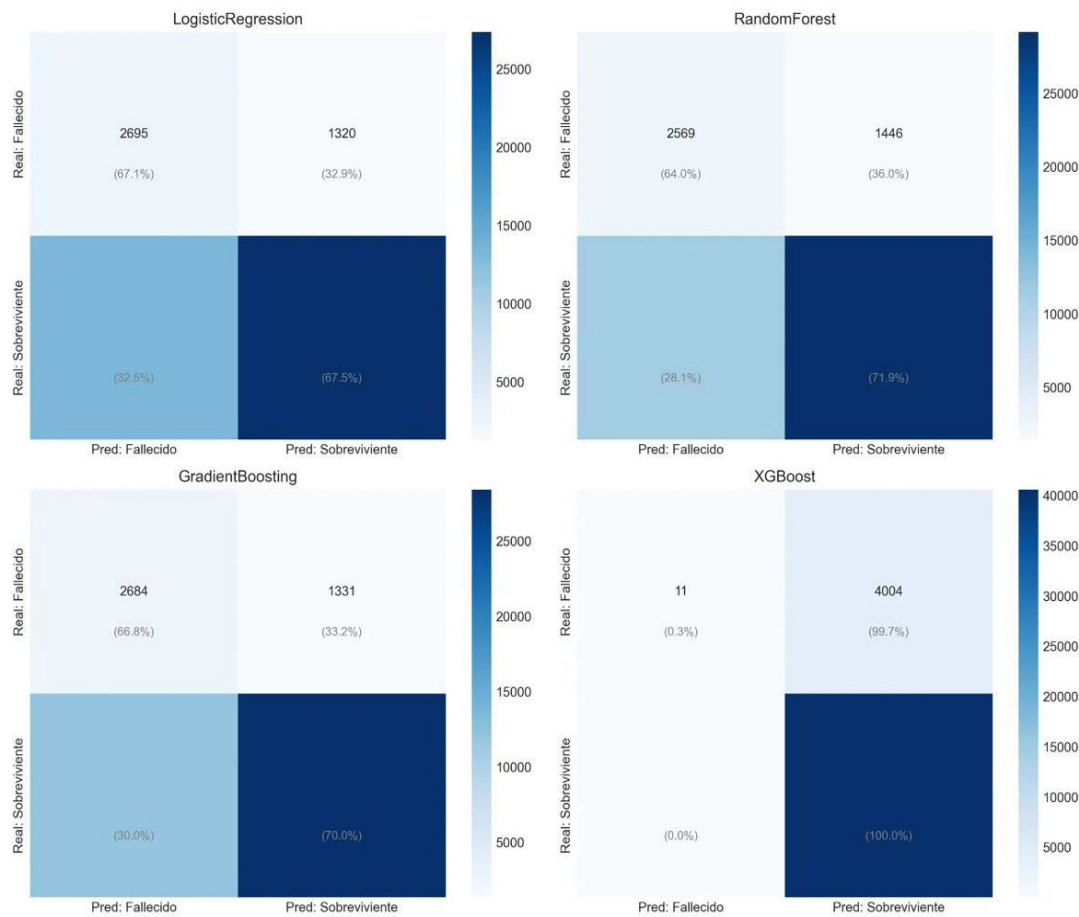


Figura 10. Matrices de confusión de los modelos en el conjunto de prueba

Fuente: Elaboración propia

Capítulo 5. DISCUSIÓN

5.1 Principales hallazgos

Los modelos de aprendizaje automático alcanzaron capacidad discriminatoria moderada pero consistente (AUC-ROC: 0.731-0.753), comparable a los nomogramas clínicos establecidos que típicamente reportan valores de 0.70-0.80 [5-8,15,16]. Gradient Boosting emergió como el modelo más balanceado, superando marginalmente a Random Forest y Regresión Logística.

La mejora del ML sobre la regresión logística fue modesta (AUC: 0.750 vs 0.731, diferencia de ~ 0.02), lo cual es consistente con la revisión de Christodoulou et al. (2019) que no encontró ventaja consistente del ML sobre métodos tradicionales en 71 estudios clínicos [24].

La confirmación de la tríada pronóstica clásica (grado tumoral, PSA, edad) como las variables más predictivas valida la coherencia clínica de los modelos y su alineamiento con décadas de investigación oncológica [4,11,12].

5.2 Comparación con la literatura

Los resultados (AUC=0.75) se sitúan en el rango medio-alto reportado en estudios recientes de ML en cáncer de próstata:

- Wong et al. (2019) reportaron AUC de 0.72 para predicción de recurrencia bioquímica con Random Forest [21]
- Agrawal et al. (2023), en una revisión sistemática, encontraron un AUC promedio de *0.78 para modelos de supervivencia*
- Los nomogramas de Kattan típicamente alcanzan C-index de 0.70-0.80 en validaciones externas [5,7]

Esto sugiere que nuestros modelos ofrecen rendimiento comparable a herramientas clínicas establecidas, aunque sin validación externa no podemos afirmar superioridad.

5.3 Limitaciones del estudio

Se identificaron las siguientes limitaciones importantes:

1. **Outcome no específico:** El 73% de las muertes fueron por causas no relacionadas con cáncer de próstata, lo que diluye la especificidad de la predicción oncológica debido a la mortalidad competitiva propia de una cohorte con mediana de edad de 65 años. El uso de supervivencia cáncer-específica como outcome alternativo habría sido más apropiado pero presenta sus propios sesgos (clasificación de causa de muerte).
2. **Ausencia de estadificación TNM completa:** La variable TNM no estaba disponible de forma consistente en la extracción realizada, limitando la incorporación de información de estadio al modelo.
3. **Desbalance de clases:** El ratio 10:1 entre supervivientes y fallecidos (91% vs 9%) presiona la sensibilidad de los clasificadores, que tienden a optimizar la métrica global a costa de subdetectar eventos minoritarios.
4. **Ausencia de validación externa:** Los resultados requieren confirmación en cohortes independientes antes de cualquier aplicación clínica. La validación interna, aunque robusta (5-fold CV), no garantiza generalizabilidad.
5. **Población predominantemente caucásica:** El 78.3% de la cohorte es de raza blanca, limitando la generalizabilidad a poblaciones con diferente composición étnica, particularmente en Latinoamérica.
6. **Falta de variables de comorbilidad:** SEER no incluye índice de Charlson u otras medidas de comorbilidad, factor importante dado que el 73% de muertes fueron por causas no oncológicas.
7. **Posible data leakage temporal:** Pacientes de años más recientes tienen menos tiempo de seguimiento, potencialmente introduciendo sesgo en la evaluación.

5.4 Implicaciones clínicas

Los modelos desarrollados NO están listos para implementación clínica directa. Antes de cualquier uso en práctica clínica se requiere:

- Validación externa en cohortes independientes, preferiblemente prospectivas y multicéntricas

- Recalibración de probabilidades para la población objetivo
- Análisis de utilidad clínica (decision curve analysis)
- Evaluación de impacto en la toma de decisiones y outcomes de pacientes

El valor actual de este trabajo radica en:

- Confirmación de la importancia de variables pronósticas establecidas
- Demostración de metodología reproducible para desarrollo de modelos predictivos
- Identificación de limitaciones metodológicas para estudios futuros
- Base de referencia para comparación con otros modelos

5.5 Fortalezas del estudio

- **Tamaño muestral robusto:** 222,876 pacientes con 20,073 eventos, excediendo ampliamente los requerimientos mínimos para desarrollo de modelos predictivos según las recomendaciones de validación TRIPOD [9]
- **Validación interna estructurada:** Partición estratificada y validación cruzada de 5 pliegues reduce riesgo de sobreajuste
- **Datos de alta calidad:** Registro SEER con tasa de completitud superior al 98% y verificación independiente de datos
- **Transparencia metodológica:** Documentación siguiendo guías TRIPOD [26]
- **Comparación de múltiples algoritmos:** Evaluación sistemática de cuatro enfoques permite conclusiones más robustas

Análisis de calibración: Evaluación de Brier score y curvas de calibración, frecuentemente omitido en estudios de ML

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

1. Los modelos de aprendizaje automático demostraron **capacidad discriminatoria moderada pero consistente** (AUC-ROC: 0.731-0.753) para predicción de supervivencia a 5 años en cáncer de próstata, comparable a nomogramas clínicos establecidos.
2. **Gradient Boosting** alcanzó el mejor balance entre discriminación (AUC=0.750), sensibilidad (70%) y especificidad (67%), siendo el algoritmo recomendado para futuras investigaciones.
3. La **tríada pronóstica clásica** (grado tumoral, PSA, edad) fue confirmada como el conjunto de variables más predictivas, validando la coherencia clínica de los modelos con décadas de investigación oncológica.
4. El uso de **supervivencia global como outcome** impone limitaciones sustanciales: el 73% de las muertes fueron por causas no oncológicas, diluyendo la especificidad de la predicción para mortalidad por cáncer de próstata.
5. Las limitaciones identificadas (ausencia de validación externa, outcome no específico, población predominantemente caucásica) **impiden la implementación clínica directa** sin estudios adicionales de validación.

El estudio proporciona una **base metodológica reproducible y transparente** para futuras investigaciones en predicción pronóstica oncológica.

6.2 Conclusiones personales

El desarrollo de este Trabajo Fin de Máster ha representado una experiencia formativa que permitió aplicar los conocimientos adquiridos durante el Máster en Análisis de Datos Masivos a un problema clínico real con relevancia personal directa.

El diagnóstico de cáncer de próstata en un familiar cercano otorgó un significado especial a esta investigación, y la transformó de un ejercicio académico a una investigación personal sobre las herramientas disponibles para comprender y predecir la evolución de esta enfermedad. Esta conexión personal proporcionó motivación sostenida durante los meses de trabajo y una perspectiva única sobre las implicaciones humanas de las predicciones estadísticas.

A nivel técnico, el proyecto permitió profundizar en:

- Manejo de grandes volúmenes de datos clínicos (>200,000 registros)
- Implementación de pipelines de machine learning robustos y reproducibles
- Interpretación crítica de resultados en el contexto de la literatura científica
- Comprensión de las limitaciones inherentes a los estudios retrospectivos

La importancia de la validación externa y las limitaciones de los datos retrospectivos emergen como lecciones fundamentales para cualquier trabajo futuro en ciencia de datos aplicada a la salud. La complejidad del ML no siempre se traduce en mejor rendimiento predictivo, y la interpretabilidad clínica debe priorizarse sobre la optimización técnica.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

7.1 Líneas de trabajo a corto plazo

- Reextraer datos SEER incluyendo estadificación TNM completa como variable predictora adicional
- Implementar análisis de supervivencia cáncer-específica como outcome alternativo para mejorar la especificidad de la predicción
- Aplicar técnicas de explicabilidad avanzadas (SHAP values, LIME) para mejorar la interpretabilidad de los modelos
- Implementar análisis de subgrupos por raza/etnicidad para evaluar equidad algorítmica

7.2 Líneas de trabajo a mediano plazo

- Validar los modelos en cohortes externas independientes (bases de datos europeas, registros hospitalarios)
- Desarrollar modelos de supervivencia basados en tiempo hasta evento (Random Survival Forest, DeepSurv) en lugar de clasificación binaria
- Integrar biomarcadores moleculares (Decipher, Oncotype DX) cuando estén disponibles en bases poblacionales
- Comparar formalmente con nomogramas clínicos establecidos (MSKCC, CAPRA) en las mismas cohortes
- Implementar Decision Curve Analysis para evaluar utilidad clínica

7.3 Líneas de trabajo a largo plazo

- Diseñar estudio prospectivo de validación en población colombiana para evaluar transportabilidad de modelos USA→Latinoamérica

- Desarrollar herramienta web interactiva para cálculo de riesgo individual (previa validación clínica rigurosa)
- Explorar integración con datos de imágenes (resonancia magnética multiparamétrica) mediante deep learning
- Evaluar impacto clínico de la implementación de modelos predictivos en toma de decisiones terapéuticas mediante ensayo clínico
- Desarrollar modelos específicos para poblaciones latinas y afrodescendientes que permitan predicción más equitativa.

Capítulo 8. REFERENCIAS

- [1] H. Sung, J. Ferlay, R.L. Siegel, et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209-249, 2021.
- [2] F. Bray, M. Laversanne, H. Sung, et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229-263, 2024.
- [3] R.L. Siegel, A.N. Giaquinto, A. Jemal, "Cancer statistics, 2024," *CA Cancer J. Clin.*, vol. 74, no. 1, pp. 12-49, 2024.
- [4] P. Rawla, "Epidemiology of Prostate Cancer," *World J. Oncol.*, vol. 10, no. 2, pp. 63-89, 2019.
- [5] M.W. Kattan, J.A. Eastham, A.M. Stapleton, T.M. Wheeler, P.T. Scardino, "A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer," *J. Natl. Cancer Inst.*, vol. 90, no. 10, pp. 766-771, 1998.
- [6] M.R. Cooperberg, D.J. Pasta, E.P. Elkin, et al., "The University of California, San Francisco Cancer of the Prostate Risk Assessment score," *J. Urol.*, vol. 173, no. 6, pp. 1938-1942, 2005.
- [7] A.J. Stephenson, P.T. Scardino, J.A. Eastham, et al., "Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy," *J. Clin. Oncol.*, vol. 23, no. 28, pp. 7005-7012, 2005.
- [8] A. Briganti, A. Larcher, F. Abdollah, et al., "Updated nomogram predicting lymph node invasion in patients with prostate cancer," *Eur. Urol.*, vol. 61, no. 3, pp. 480-487, 2012.
- [9] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement," *Ann. Intern. Med.*, vol. 162, no. 1, pp. 55-63, 2015.
- [10] K.G.M. Moons, R.F. Wolff, R.D. Riley, et al., "PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51-58, 2019.
- [11] J.I. Epstein, L. Egevad, M.B. Amin, et al., "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma," *Am. J. Surg. Pathol.*, vol. 40, no. 2, pp. 244-252, 2016.

- [12] J.I. Epstein, M.J. Zelefsky, D.D. Sjoberg, et al., "A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score," *Eur. Urol.*, vol. 69, no. 3, pp. 428-435, 2016.
- [13] W.J. Catalona, D.S. Smith, T.L. Ratliff, et al., "Measurement of prostate-specific antigen in serum as a screening test for prostate cancer," *N. Engl. J. Med.*, vol. 324, no. 17, pp. 1156-1161, 1991.
- [14] M.B. Amin, F.L. Greene, S.B. Edge, et al., "The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more personalized approach to cancer staging," *CA Cancer J. Clin.*, vol. 67, no. 2, pp. 93-99, 2017.
- [15] J.J. Tosoian, M. Chappidi, Z. Feng, et al., "Prediction of pathological stage based on clinical stage, serum PSA, and biopsy Gleason score: Partin Tables in the contemporary era," *BJU Int.*, vol. 119, no. 5, pp. 676-683, 2017.
- [16] S.F. Shariat, P.I. Karakiewicz, C.G. Roehrborn, M.W. Kattan, "An updated catalog of prostate cancer predictive tools," *Cancer*, vol. 113, no. 11, pp. 3075-3099, 2008.
- [17] A. Rajkomar, J. Dean, I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347-1358, 2019.
- [18] E.J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44-56, 2019.
- [19] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H.J.W.L. Aerts, "Artificial intelligence in radiology," *Nat. Rev. Cancer*, vol. 18, no. 8, pp. 500-510, 2018.
- [20] S.L. Goldenberg, G. Nir, S.E. Salcudean, "A new era: artificial intelligence and machine learning in prostate cancer," *Nat. Rev. Urol.*, vol. 16, no. 7, pp. 391-403, 2019.
- [21] N.C. Wong, C. Lam, L. Patterson, B. Shayegan, "Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy," *BJU Int.*, vol. 123, no. 1, pp. 51-57, 2019.
- [22] A.J. Hung, J. Chen, I.S. Gill, "Automated Performance Metrics and Machine Learning Algorithms to Measure Surgeon Performance in Robotic Surgery," *JAMA Surg.*, vol. 153, no. 8, pp. 770-771, 2018.
- [23] K.A. Tran, O. Kondrashova, A. Bradley, E.D. Williams, J.V. Pearson, N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," *Genome Med.*, vol. 13, no. 1, p. 152, 2021.
- [24] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, "A systematic

review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12-22, 2019.

- [25] M. Nagendran, Y. Chen, C.A. Lovejoy, et al., "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, p. m689, 2020.
- [26] K.G.M. Moons, J.A.H. de Groot, W. Bouwmeester, et al., "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist," *PLoS Med.*, vol. 11, no. 10, e1001744, 2014.
- [27] R.F. Wolff, K.G.M. Moons, R.D. Riley, et al., "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration," *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51-58, 2019.
- [28] National Cancer Institute, "Surveillance, Epidemiology, and End Results (SEER) Program," SEER*Stat Database, 2024. [Online]. Available: <https://seer.cancer.gov/>
- [29] J.B. Yu, C.P. Gross, L.D. Wilson, B.D. Smith, "NCI SEER public-use data: applications and limitations in oncology research," *Oncology (Williston Park)*, vol. 23, no. 3, pp. 288-295, 2009

Capítulo 9. ANEXOS

9.1.1 Anexo A: Código de preprocesamiento de datos

A continuación, se presenta un fragmento representativo del script `preparar_datos_seer.py` utilizado para la limpieza y transformación de datos:

Fragmento A.1: Criterios de exclusión de la cohorte

```
# =====
# PASO 3: APLICAR CRITERIOS DE EXCLUSION
# =====
print("\n" + "="*70)
print("PASO 3: APLICANDO CRITERIOS DE EXCLUSION")
print("="*70)

n_inicial = len(df)
print(f"Casos iniciales: {n_inicial:,}")

# Criterio 1: Solo tratamiento curativo (cirugía y/o radioterapia)
df = df[df['tiene_tratamiento_curativo'] == True].copy()
n_post_trat = len(df)
print(f"Tras filtro tratamiento curativo: {n_post_trat:,} "
      f"(excluidos: {n_inicial - n_post_trat:,})")

# Criterio 2: PSA documentado
df = df[df['psa_valido'] == True].copy()
n_post_psa = len(df)
print(f"Tras filtro PSA documentado: {n_post_psa:,} "
      f"(excluidos: {n_post_trat - n_post_psa:,})")

# Criterio 3: Grado tumoral documentado
df = df[df['grado_valido'] == True].copy()
n_post_grado = len(df)
print(f"Tras filtro grado documentado: {n_post_grado:,} "
      f"(excluidos: {n_post_psa - n_post_grado:,})")

# Criterio 4: Seguimiento adecuado
df = df[(df['meses_supervivencia'] >= 60) |
        (df['estado_vital'] == 'Dead')].copy()
n_final = len(df)
print(f"Tras filtro seguimiento: {n_final:,} "
      f"(excluidos: {n_post_grado - n_final:,})")

print(f"\nCOHORTE FINAL: {n_final:,} casos "
      f"({n_final/n_inicial*100:.1f}% del total)")
```

Fragmento A.2: Creación de la variable outcome (supervivencia a 5 años)

```
def clasificar_supervivencia(row):
    """
    Clasifica supervivencia a 5 años (60 meses).
    Returns: 1 si sobrevivió >= 60 meses, 0 si falleció antes
    """
    if row['meses_supervivencia'] >= 60:
        return 1 # Sobrevivió al menos 5 años
    elif row['estado_vital'] == 'Dead':
        return 0 # Falleció antes de 5 años
    else:
        return None # Censurado antes de 60 meses (excluir)
```

```
df['supervivencia_5_anos'] = df.apply(clasificar_supervivencia, axis=1)
df = df.dropna(subset=['supervivencia_5_anos'])
df['supervivencia_5_anos'] = df['supervivencia_5_anos'].astype(int)
```

Fragmento A.3: Preprocesamiento de variables clínicas

```
def convertir_psa(valor):
    """Convierte PSA a valor numerico, manejando rangos SEER."""
    if pd.isna(valor) or valor in ['Blank(s)', 'Unknown']:
        return np.nan
    valor_str = str(valor).strip()
    # Patron numerico directo
    try:
        return float(valor_str)
    except ValueError:
        pass
    # Rangos como "0.1-3.9"
    match = re.match(r'(\d+\.\d*)\s*[\-]\s*(\d+\.\d*)', valor_str)
    if match:
        return (float(match.group(1)) + float(match.group(2))) / 2
    return np.nan

def categorizar_psa(psa):
    """Categoriza PSA segun umbrales clinicos estandar."""
    if pd.isna(psa):
        return 'Desconocido'
    if psa < 4:
        return '<4 ng/mL'
    elif psa < 10:
        return '4-10 ng/mL'
    elif psa < 20:
        return '10-20 ng/mL'
    else:
        return '>20 ng/mL'

def convertir_grado(valor):
    """Convierte grado tumoral SEER a numerico (1-4)."""
    mapa = {
        'Well differentiated; Grade I': 1,
        'Moderately differentiated; Grade II': 2,
        'Poorly differentiated; Grade III': 3,
        'Undifferentiated; anaplastic; Grade IV': 4
    }
    return mapa.get(valor, np.nan)

def categorizar_tratamiento(row):
    """Clasifica modalidad de tratamiento."""
    cirugia = row.get('cirugia_realizada', False)
    radio = row.get('radioterapia_realizada', False)
    if cirugia and radio:
        return 'Cirugia + Radioterapia'
    elif cirugia:
        return 'Solo Cirugia'
    elif radio:
        return 'Solo Radioterapia'
    else:
        return 'Sin tratamiento curativo'

# Aplicar transformaciones
df['psa_numerico'] = df['psa_raw'].apply(convertir_psa)
df['grado_numerico'] = df['grado_raw'].apply(convertir_grado)
df['tipo_tratamiento'] = df.apply(categorizar_tratamiento, axis=1)
```

Anexo B: Código de entrenamiento de modelos

Fragmento del script `entrenar_modelos_ml.py` con la implementación de los algoritmos:

Fragmento B.1: Configuración del pipeline de preprocesamiento

```
import numpy as np
import pandas as pd
from sklearn.model_selection import StratifiedKFold, train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier
from sklearn.metrics import (roc_auc_score, accuracy_score, f1_score,
                             precision_score, recall_score, confusion_matrix,
                             classification_report, roc_curve, brier_score_loss)

RANDOM_STATE = 42

# Division estratificada 80/20
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=RANDOM_STATE, stratify=y
)

# Feature engineering: transformacion Logaritmica de PSA
X_train['psa_log'] = np.log1p(X_train['psa_numerico'])
X_test['psa_log'] = np.log1p(X_test['psa_numerico'])

# Escalado de variables numericas
scaler = StandardScaler()
cols_numericas = ['edad_numerica', 'psa_numerico', 'psa_log']
X_train[cols_numericas] = scaler.fit_transform(X_train[cols_numericas])
X_test[cols_numericas] = scaler.transform(X_test[cols_numericas])
```

Fragmento B.2: Definición y entrenamiento de los 4 modelos

```
# Definicion de modelos con hiperparametros optimizados
modelos = {
    'Logistic Regression': LogisticRegression(
        C=1.0,
        class_weight='balanced',
        max_iter=1000,
        random_state=RANDOM_STATE,
        solver='lbfgs'
    ),
    'Random Forest': RandomForestClassifier(
        n_estimators=200,
        max_depth=10,
        min_samples_split=20,
        min_samples_leaf=10,
        class_weight='balanced',
        random_state=RANDOM_STATE,
        n_jobs=-1
    ),
    'Gradient Boosting': GradientBoostingClassifier(
        n_estimators=200,
        max_depth=4,
        learning_rate=0.1,
        min_samples_split=20,
        min_samples_leaf=10,
        subsample=0.8,
        random_state=RANDOM_STATE
    ),
}
```

```
'XGBoost': XGBClassifier(  
    n_estimators=200,  
    max_depth=4,  
    learning_rate=0.1,  
    min_child_weight=10,  
    subsample=0.8,  
    colsample_bytree=0.8,  
    scale_pos_weight=ratio_clases,  
    random_state=RANDOM_STATE,  
    use_label_encoder=False,  
    eval_metric='logloss'  
)  
}
```

Fragmento B.3: Validación cruzada estratificada y evaluación

```
# Validacion cruzada estratificada 5-fold  
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=RANDOM_STATE)  
  
resultados_cv = {}  
for nombre, modelo in modelos.items():  
    print(f"\nEntrenando {nombre}...")  
    metricas_folds = {  
        'auc': [], 'accuracy': [], 'f1': [],  
        'precision': [], 'recall': [], 'brier': []  
    }  
  
    for fold, (idx_train, idx_val) in enumerate(cv.split(X_train, y_train)):  
        X_fold_train = X_train.iloc[idx_train]  
        X_fold_val = X_train.iloc[idx_val]  
        y_fold_train = y_train.iloc[idx_train]  
        y_fold_val = y_train.iloc[idx_val]  
  
        modelo_fold = clone(modelo)  
        modelo_fold.fit(X_fold_train, y_fold_train)  
  
        y_pred = modelo_fold.predict(X_fold_val)  
        y_prob = modelo_fold.predict_proba(X_fold_val)[:, 1]  
  
        metricas_folds['auc'].append(roc_auc_score(y_fold_val, y_prob))  
        metricas_folds['accuracy'].append(accuracy_score(y_fold_val, y_pred))  
        metricas_folds['f1'].append(f1_score(y_fold_val, y_pred))  
        metricas_folds['precision'].append(precision_score(y_fold_val, y_pred))  
        metricas_folds['recall'].append(recall_score(y_fold_val, y_pred))  
        metricas_folds['brier'].append(brier_score_loss(y_fold_val, y_prob))  
  
    resultados_cv[nombre] = {  
        'metrica': {'media': np.mean(vals), 'std': np.std(vals)}  
        for metrica, vals in metricas_folds.items()  
    }  
  
# Evaluacion final en conjunto de test  
for nombre, modelo in modelos.items():  
    modelo.fit(X_train, y_train)  
    y_pred_test = modelo.predict(X_test)  
    y_prob_test = modelo.predict_proba(X_test)[:, 1]  
    print(f"\n{nOMBRE} - Test AUC: "  
        f"{roc_auc_score(y_test, y_prob_test):.4f}")
```

Anexo C: Análisis estadístico y visualizaciones

Fragmentos del script generar_visualizaciones.py con el análisis de supervivencia y los tests estadísticos aplicados:

Fragmento C.1: Curvas de Kaplan-Meier estratificadas por grado tumoral

```
from lifelines import KaplanMeierFitter
from lifelines.statistics import multivariate_logrank_test

# Kaplan-Meier por Grado Tumoral
fig, ax = plt.subplots(figsize=(10, 8))
kmf = KaplanMeierFitter()
grado_labels = ['Grado I (Bien dif.)', 'Grado II (Mod. dif.)',
                'Grado III (Pobre dif.)', 'Grado IV (Indif.)']

for i, grado in enumerate([1, 2, 3, 4]):
    mask = df['grado_numerico'] == grado
    if mask.sum() > 0:
        tiempo = np.minimum(df.loc[mask, 'meses_supervivencia'], 60)
        evento = df.loc[mask, 'evento']
        kmf.fit(tiempo, evento,
                label=f'{grado_labels[i]} (n={mask.sum():,})')
        kmf.plot_survival_function(ax=ax, ci_show=True,
                                   color=colores_grado[i], linewidth=2)

# Log-rank test
resultado_logrank = multivariate_logrank_test(
    np.minimum(df['meses_supervivencia'], 60),
    df['grado_numerico'],
    df['evento']
)

ax.set_xlabel('Tiempo (meses)')
ax.set_ylabel('Probabilidad de Supervivencia')
ax.set_title('Curvas de Kaplan-Meier por Grado Tumoral')
ax.set_xlim(0, 60)
ax.set_ylim(0.7, 1.0)
ax.legend(loc='Lower Left')
```

Fragmento C.2: Tests estadísticos de comparación entre grupos

```
from scipy import stats

# Test Mann-Whitney U para variables continuas
stat_edad, p_edad = stats.mannwhitneyu(
    df_sobrev['edad_numerica'],
    df_fallec['edad_numerica']
)

stat_psa, p_psa = stats.mannwhitneyu(
    df_sobrev['psa_numerico'],
    df_fallec['psa_numerico']
)

# Test Chi-cuadrado para variables categoricas
contingencia_grado = pd.crosstab(
    df['supervivencia_5_anos'], df['grado_numerico']
)
chi2_grado, p_chi2_grado, dof_grado, _ = stats.chi2_contingency(
    contingencia_grado
)

contingencia_psa = pd.crosstab(
    df['supervivencia_5_anos'], df['psa_cat']
)
chi2_psa, p_chi2_psa, dof_psa, _ = stats.chi2_contingency(
    contingencia_psa
)

print(f"Mann-Whitney U (Edad): U={stat_edad:.2e}, p<0.001")
```

Anatoly Cerchiaro Peña

```
print(f"Mann-Whitney U (PSA): U={stat_psa:.2e}, p<0.001")
print(f"Chi-cuadrado (Grado): X2={chi2_grado:.1f}, "
      f"gl={dof_grado}, p<0.001")
print(f"Chi-cuadrado (PSA): X2={chi2_psa:.1f}, "
      f"gl={dof_psa}, p<0.001")
```

Datos SEER: Los datos del programa SEER son de acceso público para investigadores mediante solicitud formal al National Cancer Institute. El proceso de solicitud puede iniciarse en: [https:// seer.cancer.gov/data/access.html](https://seer.cancer.gov/data/access.html)

Reproducibilidad: Para garantizar la reproducibilidad de los resultados, se utilizó semilla aleatoria fija (random_state=42) en todas las operaciones que involucran aleatorización

[PÁGINA INTENCIONADAMENTE EN BLANCO]