Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: Medicina

Detección de Metástasis en Cáncer de Mama y Cáncer de Próstata mediante Machine Learning a partir de RNA-seq

---

Autor: Sua de la Cruz Odriozola

Tutora: Erola Pairó Castiñeira

Profesora: Laia Subirats Maté

San Sebastián, 11 de junio de 2024

FICHA DEL TRABAJO FINAL

Título del trabajo:	Detección de Metástasis en Cáncer de Mama y Cáncer de Próstata mediante Machine Learning a partir de RNA-seq
Autor:	Sua de la Cruz Odriozola
Tutora:	Elora Pairó Castiñeira
Profesora responsable:	Laia Subirats Maté
Fecha de entrega:	06/2024
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Medicina
Idioma del trabajo:	Español
Palabras clave:	Machine learning, metástasis, RNA-seq
Código fuente:	https://github.com/suadlco/TFM.git

Abstract

Cancer is a disease that causes millions of deaths per year, with metastasis occurring in the majority of deaths. The aim of this work is to create machine learning models of metastasis from gene expression data, in order to detect whether a person has metastasis or not, evaluating their performance using various metrics and tools, and comparing their results to identify the most effective approach.

As a result, these models could not only reduce cancer- and metastasis-related mortality, but also prevent its development and facilitate the identification of critical biomarkers for early diagnosis and treatment.

Key words: Machine Learning, Breast cancer, Prostate cancer, Metastasis, Gene expression, RNA-seq.

Resumen

El cáncer es una enfermedad que causa millones de muertes al año, produciéndose metástasis en la mayoría de los fallecimientos. El objetivo de este trabajo es crear modelos de machine learning de metástasis a partir de datos de expresión génica, con la finalidad de detectar si una persona tiene metástasis o no, evaluando su rendimiento mediante diversas métricas y herramientas, y comparando sus resultados para identificar el enfoque más eficaz.

En consecuencia, estos modelos podrían no solo reducir la mortalidad relacionada con el cáncer y la metástasis, sino también prevenir su desarrollo y facilitar la identificación de biomarcadores críticos para el tratamiento y diagnóstico temprano.

Palabras clave: Machine Learning, Cáncer de mama, Cáncer de próstata, Metástasis, Expresión génica, RNA-seq.

Índice general

Abstract v

Resumen vii

Índice ix

Índice de Figuras xiii

Índice de Tablas xv

Índice de Programas xvii

1. Introducción 1

1.1. Descripción general del problema . . . 1
1.2. Motivación personal . . . 1
1.3. Objetivos . . . 2
1.3.1. Objetivo Principal . . . 2
1.3.2. Objetivos Secundarios . . . 2
1.4. Planificación . . . 3
1.4.1. Entregables . . . 3
1.4.2. EDT . . . 3
1.4.3. Gestión del tiempo . . . 5

2. Estado del Arte 7

2.1. Introducción . . . 7
2.2. Justificación del Tema de Investigación . . . 7
2.3. Metodología . . . 8
2.4. Conclusiones . . . 10

3. Tecnologías y métodos **11**
3.1. Tecnologías usadas en el proyecto . . . 11
3.1.1. R . . . 11
3.1.2. Python . . . 12
3.2. Tecnologías para el desarrollo . . . 13
3.2.1. RStudio . . . 13
3.2.2. Google Colab . . . 13
3.2.3. Google Drive . . . 13
3.3. Modelos Machine Learning . . . 14
3.3.1. Naive Bayes . . . 14
3.3.2. SVC . . . 14
3.3.3. Regresión Logística . . . 15
3.3.4. Random Forest . . . 15
3.3.5. XGBoost . . . 16
3.3.6. Deep Neural Network . . . 16
3.3.7. TabNet . . . 16
3.3.8. AutoGluon: TabularPredictor . . . 17
3.4. Métricas y herramientas de evaluación . . . 17
3.4.1. AUC-ROC . . . 17
3.4.2. Matriz de confusión . . . 17
3.4.3. Precisión, Recall y F1-score . . . 18

4. Procesamiento de datos **19**
4.1. Sobre el dataset . . . 19
4.2. Construcción del dataset . . . 21
4.3. Conjunto de datos . . . 22
4.3.1. Conjunto de validación . . . 24
4.3.2. Conjunto de entrenamiento . . . 24

5. Resultados **27**
5.1. Naive Bayes . . . 27
5.1.1. Evaluación del modelo . . . 28
5.2. SVC . . . 31
5.2.1. Evaluación del modelo . . . 31
5.3. Regresión logística . . . 34
5.3.1. Evaluación del modelo . . . 35
5.4. Random Forest . . . 38

5.4.1. Evaluación del modelo 38
5.5. Gradient boosting: XGBoost 41
5.5.1. Evaluación del modelo 42
5.6. Deep Neural Network 45
5.6.1. Evaluación del modelo 45
5.7. TabNetClassifier 48
5.7.1. Evaluación del modelo 49
5.8. AutoGluon: TabularPredictor 52
5.8.1. Evaluación del modelo 52
5.9. Análisis de influencia de genes 55

6. Conclusiones y trabajos futuros **59**

A. Código para la creación del dataset **61**
A.1. Extracción de datos 61
A.2. Limpieza y normalización de datos 63
A.3. Creación de datos sintéticos mediante GAN 64

B. Funciones auxiliares para la evaluación de los modelos **67**

C. Código del análisis de los genes **69**

Bibliografía **70**

Índice de figuras

1.1. LDE diagrama 5
1.2. Diagrama de Gantt 6

5.1. Resultados del modelo NB-ST para cáncer de próstata 28
5.2. Resultados del modelo NB-ST para cáncer de mama 28
5.3. Resultados del modelo NB-ADA para cáncer de próstata 29
5.4. Resultados del modelo NB-ADA para cáncer de mama 29
5.5. Resultados del modelo NB-GAN para cáncer de próstata 30
5.6. Resultados del modelo NB-GAN para cáncer de mama 30
5.7. Resultados del modelo SVC-ST para cáncer de próstata 31
5.8. Resultados del modelo SVC-ST para cáncer de mama 32
5.9. Resultados del modelo SVC-ADA para cáncer de próstata 32
5.10. Resultados del modelo SVC-ADA para cáncer de mama 33
5.11. Resultados del modelo SVC-GAN para cáncer de próstata 33
5.12. Resultados del modelo SVC-GAN para cáncer de mama 34
5.13. Resultados del modelo LR-ST para cáncer de próstata 35
5.14. Resultados del modelo LR-ST para cáncer de mama 35
5.15. Resultados del modelo LR-ADA para cáncer de próstata 36
5.16. Resultados del modelo LR-ADA para cáncer de mama 36
5.17. Resultados del modelo LR-GAN para cáncer de próstata 37
5.18. Resultados del modelo LR-GAN para cáncer de mama 37
5.19. Resultados del modelo RF-ST para cáncer de próstata 39
5.20. Resultados del modelo RF-ST para cáncer de mama 39
5.21. Resultados del modelo RF-ADA para cáncer de próstata 40
5.22. Resultados del modelo RF-ADA para cáncer de mama 40
5.23. Resultados del modelo RF-GAN para cáncer de próstata 41
5.24. Resultados del modelo RF-GAN para cáncer de mama 41
5.25. Resultados del modelo XGB-ST para cáncer de próstata 42

5.26. Resultados del modelo XGB-ST para cáncer de mama 43
5.27. Resultados del modelo XGB-ADA para cáncer de próstata 43
5.28. Resultados del modelo XGB-ADA para cáncer de mama 44
5.29. Resultados del modelo XGB-GAN para cáncer de próstata 44
5.30. Resultados del modelo XGB-GAN para cáncer de mama 45
5.31. Resultados del modelo DNN-ST para cáncer de próstata 46
5.32. Resultados del modelo DNN-ST para cáncer de mama 46
5.33. Resultados del modelo DNN-ADA para cáncer de próstata 47
5.34. Resultados del modelo DNN-ADA para cáncer de mama 47
5.35. Resultados del modelo DNN-GAN para cáncer de próstata 48
5.36. Resultados del modelo DNN-GAN para cáncer de mama 48
5.37. Resultados del modelo TABNET-ST para cáncer de próstata 49
5.38. Resultados del modelo TABNET-ST para cáncer de mama 49
5.39. Resultados del modelo TABNET-ADA para cáncer de próstata 50
5.40. Resultados del modelo TABNET-ADA para cáncer de mama 50
5.41. Resultados del modelo TABNET-GAN para cáncer de próstata 51
5.42. Resultados del modelo TABNET-GAN para cáncer de mama 51
5.43. Resultados del modelo TP-ST para cáncer de próstata 52
5.44. Resultados del modelo TP-ST para cáncer de mama 53
5.45. Resultados del modelo TP-ADA para cáncer de próstata 53
5.46. Resultados del modelo TP-ADA para cáncer de mama 54
5.47. Resultados del modelo TP-GAN para cáncer de próstata 54
5.48. Resultados del modelo TP-GAN para cáncer de mama 55
5.49. Genes que más influyen en el cáncer de próstata 56
5.50. Genes que más influyen en el cáncer de mama 57
5.51. Genes con coeficiente positivo que más influyen en el cáncer de próstata 57
5.52. Genes con coeficiente positivo que más influyen en el cáncer de mama 58

Índice de tablas

1.1. Tabla de los entregables 3

4.1. Casos disponibles de metástasis por tipo de cáncer 19
4.2. Cantidad de casos de cáncer de próstata 20
4.3. Cantidad de casos de cáncer de mama 20
4.4. Dimensiones de los archivos 22
4.5. Dimensiones de los datos 22
4.6. Proporciones de las clases para cada tipo de cáncer 23
4.7. Cantidad de casos por conjunto 24
4.8. Cantidad de casos por método 25

5.1. Métricas del modelo NB-ST para cáncer de próstata 28
5.2. Métricas del modelo NB-ST para cáncer de mama 29
5.3. Métricas del modelo NB-ADA para cáncer de próstata 29
5.4. Métricas del modelo NB-ADA para cáncer de mama 30
5.5. Métricas del modelo NB-GAN para cáncer de próstata 30
5.6. Métricas del modelo NB-GAN para cáncer de mama 31
5.7. Métricas del modelo SVC-ST para cáncer de próstata 32
5.8. Métricas del modelo SVC-ST para cáncer de mama 32
5.9. Métricas del modelo SVC-ADA para cáncer de próstata 33
5.10. Métricas del modelo SVC-ADA para cáncer de mama 33
5.11. Métricas del modelo SVC-GAN para cáncer de próstata 34
5.12. Métricas del modelo SVC-GAN para cáncer de mama 34
5.13. Métricas del modelo LR-ST para cáncer de próstata 35
5.14. Métricas del modelo LR-ST para cáncer de mama 36
5.15. Métricas del modelo LR-ADA para cáncer de próstata 36
5.16. Métricas del modelo LR-ADA para cáncer de mama 37
5.17. Métricas del modelo LR-GAN para cáncer de próstata 37

5.18. Métricas del modelo LR-GAN para cáncer de mama 38
5.19. Métricas del modelo RF-ST para cáncer de próstata 38
5.20. Métricas del modelo RF-ST para cáncer de mama 39
5.21. Métricas del modelo RF-ADA para cáncer de próstata 39
5.22. Métricas del modelo RF-ADA para cáncer de mama 40
5.23. Métricas del modelo RF-GAN para cáncer de próstata 40
5.24. Métricas del modelo RF-GAN para cáncer de mama 41
5.25. Métricas del modelo XGB-ST para cáncer de próstata 42
5.26. Métricas del modelo XGB-ST para cáncer de mama 42
5.27. Métricas del modelo XGB-ADA para cáncer de próstata 43
5.28. Métricas del modelo XGB-ADA para cáncer de mama 43
5.29. Métricas del modelo XGB-GAN para cáncer de próstata 44
5.30. Métricas del modelo XGB-GAN para cáncer de mama 44
5.31. Métricas del modelo DNN-ST para cáncer de próstata 45
5.32. Métricas del modelo DNN-ST para cáncer de mama 46
5.33. Métricas del modelo DNN-ADA para cáncer de próstata 46
5.34. Métricas del modelo DNN-ADA para cáncer de mama 47
5.35. Métricas del modelo DNN-GAN para cáncer de próstata 47
5.36. Métricas del modelo DNN-GAN para cáncer de mama 48
5.37. Métricas del modelo TABNET-ST para cáncer de próstata 49
5.38. Métricas del modelo TABNET-ST para cáncer de mama 50
5.39. Métricas del modelo TABNET-ADA para cáncer de próstata 50
5.40. Métricas del modelo TABNET-ADA para cáncer de mama 51
5.41. Métricas del modelo TABNET-GAN para cáncer de próstata 51
5.42. Métricas del modelo TABNET-GAN para cáncer de mama 52
5.43. Métricas del modelo TP-ST para cáncer de próstata 52
5.44. Métricas del modelo TP-ST para cáncer de mama 53
5.45. Métricas del modelo TP-ADA para cáncer de próstata 53
5.46. Métricas del modelo TP-ADA para cáncer de mama 54
5.47. Métricas del modelo TP-GAN para cáncer de próstata 54
5.48. Métricas del modelo TP-GAN para cáncer de mama 55

Índice de Programas

4.1. Función para balancear el conjunto de validación 24
4.2. Función para balancear el conjunto de validación 25
A.1. Funciones para la extracción de datos 61
A.2. Creación de los archivos para el cáncer de próstata 62
A.3. Creación de los archivos para el cáncer de mama 62
A.4. Función para filtrar y normalizar los datos 63
A.5. Entrenamiento del modelo para generar datos sintéticos 64
B.1. Función para visualizar las curvas de entrenamiento y evaluación de las redes neuronales 67
B.2. Función para mostar la matriz de confusión 67
B.3. Función para mostrar el gráfico de la cuva ROC 68
C.1. Visualización de los genes más influyentes en la detección de metástasis 69
C.2. Visualización de los genes con coeficientes positivos más influyentes en la detección de metástasis 69

Capítulo 1

Introducción

1.1. Descripción general del problema

El cáncer es una de las mayores causas de muerte de la actualidad, siendo esta la causa de muerte de millones de personas anualmente, produciéndose metástasis en el 90 % de las muertes por cáncer [1]. Por ello, la identificación precoz de esta enfermedad y su propagación a través de la metástasis representa un desafío crítico en la medicina moderna. En este Trabajo Final de Máster, se aborda el uso de técnicas de machine learning para la predicción de metástasis.

En este contexto, el análisis de la expresión génica emerge como una herramienta prometedora, permitiendo la identificación de biomarcadores y encontrar un tratamiento eficaz para el paciente [2, 3]. Por otro lado, la aplicación de técnicas de machine learning brinda la oportunidad de detectar patrones sutiles que podrían señalar la existencia de cáncer o la probabilidad de que se vaya a producir metástasis.

Estos modelos de aprendizaje automático pueden ser creados con datos clínicos de pacientes ya diagnosticados, facilitando el desarrollo de los algoritmos de detección. Además, de esta manera también es posible personalizar la terapia de los pacientes en función de sus características genéticas, pudiendo prevenir la aparición de la metástasis.

En resumen, con este Trabajo Final de Máster se pretende investigar el potencial del machine learning en el análisis de datos genéticos para mejorar la prevención de la metástasis, con el objetivo de reducir significativamente el impacto de esta enfermedad en la población mundial.

1.2. Motivación personal

La elección de este proyecto surge de una motivación personal en el deseo de contribuir a la sociedad en el ámbito de la salud. El cáncer es una enfermedad que afecta a una gran parte de la población mundial, lo que me ha impulsado a querer buscar soluciones que puedan mejorar

las tasas de detección temprana y así aplicar el tratamiento adecuado a tiempo.

Además, la aplicación de técnicas de machine learning a datos de expresión genética para la detección del cáncer y la predicción de metástasis es una oportunidad única para mejorar mis habilidades técnicas mientras genero un impacto positivo en la salud humana.

Asimismo, me inspira saber que el impacto que puede tener este proyecto en la sociedad es enorme, ya que la pedicción precisa de la metástasis no solo puede salvar vidas, sino que también puede mejorar la calidad de vida de los pacientes, así como de sus familiares y amigos, al proporcionar tratamientos más efectivos. Esta motivación me impulsa a dedicar tiempo y esfuerzo a este proyecto, con la esperanza de que los resultados obtenidos puedan contribuir significativamente a la lucha global contra el cáncer.

1.3. Objetivos

1.3.1. Objetivo Principal

El objetivo principal de este proyecto consiste en desarrollar un sistema de detección de cáncer y predicción de metástasis basado en técnicas de machine learning, utilizando datos de expresión génica con la finalidad de mejorar la precisión y eficacia en la identificación temprana de la enfermedad y su propagación.

1.3.2. Objetivos Secundarios

1. Recopilar y preparar conjuntos de datos de expresión génica relevantes para entrenar y validar modelos de machine learning.

2. Investigar y aprender sobre los datos genéticos para identificar los que puedan ser indicativos de la presencia de cáncer y la probabilidad de metástasis.

3. Explorar y aplicar diferentes algoritmos de machine learning para determinar cuáles son las más adecuadas para la detección del cáncer y la predicción de metástasis.

4. Entrenar y optimizar los modelos de machine learning con técnicas diferentes, mejorando la precisión y generalización.

5. Evaluar el rendimiento del sistema desarrollado utilizando métricas de evaluación estándar.

6. Comparar el rendimiento del sistema desarrollado con otros métodos de detección y predicción para evaluar su superioridad en términos de precisión, eficacia y eficiencia.

7. Investigar la capacidad del sistema para identificar biomarcadores genéticos asociados con la metástasis, y su utilidad en la personalización de terapias.

8. Documentar claramente los resultados obtenidos, así como las conclusiones y posibles implicaciones clínicas del sistema desarrollado, a través de la memoria del Trabajo Final de Máster.

1.4. Planificación

1.4.1. Entregables

Los entregables que van a ser presentados a la tutora del TFM son los siguientes: la definición del TFM (EDEF), el estado del arte (EEA), la implementación (EIMP), la memoria del TFM (EMEM), y la presentación audiovisual (EPA). En la tabla 1.1 aparecen los identificadores de los entregables, sus descripciones y las fechas de entrega.

Identificador	Descripción	Fecha de entrega
EDEF	Definición del TFM	12/03/2024
EEA	Estado del arte	26/03/2024
EIMP	Implementación	21/05/2024
EMEM	Memoria del TFM	11/06/2024
EPA	Presentación audiovisual	18/06/2024

Tabla 1.1: Tabla de los entregables

1.4.2. EDT

Para realizar este proyecto de manera ordenada se ha creado el diagrama de Estructura de Desglose de Trabajo (EDT). Es decir, se han definido los siguientes paquetes de trabajo: Conocimiento, Producto, Gestión y Documentación. A continuación se explican los paquetes de trabajo que aparecen en la imagen 1.1:

Conocimiento

1. **Tecnológico (T)**: este paquete abarca los pasos necesarios para adquirir el conocimiento tecnológico necesario para un buen desarrollo del proyecto. Incluye acciones como la lectura de documentación de los paquetes necesarios para la obtención de datos de caldiad y el desarrollo de los algoritmos de machine learning.

2. **Biosanitario (B)**: este paquete abarca los pasos necesarios para adquirir el conocimiento biosanitario necesario para un buen desarrollo del proyecto. Esto incluye lecturas de artículos científicos sobre temas como el cáncer, la metástasis y la información genética.

Producto

1. **Conjuntos de datos (CD)**: este paquete abarca todos los procesos necesarios para la recopilación y preparación de los conjuntos de datos genéticos y clínicos que se van a utilizar en el desarrollo de los modelos.

2. **Desarrollo de los modelos**

   a) **Entrenamiento (EN)**: abarca las tareas necesarias para el desarrollo y entrenamiento de los modelos de machine learning.

   b) **Evaluación (EV)**: incluye las tareas necesarias para evaluar el rendimiento del sistema desarrollado.

   c) **Comparación (DC)**: este paquete abarca las tareas en las que se compara el rendimiento de los modelos desarrollados.

3. **Conclusiones (C)**: recoge las tareas necesarias para interpretar los resultados y sacar conclusiones del modelo desarrollado, así como la capacidad del sistema de identificar biomarcadores genéticos y su aplicación en terapias personalizadas.

Gestión

1. **Planificación (P)**: este paquete incluye las tareas necesarias para el desarrollo de la planificación inicial.

2. **Seguimiento y Control (SC)**: incluye las tareas necesarias para garantizar el correcto desarrollo del proyecto.

3. **Reuniones (R)**: recoge las reuniones que se deben tener con el tutor para sacar adelante el proyecto.

Documentación

1. **Memoria (M)**: recoge las tareas necesarias para el desarrollo de la memoria del TFM.

2. **Defensa (D)**: incluye las tareas necesarias para llevar a cabo la defensa del proyecto.

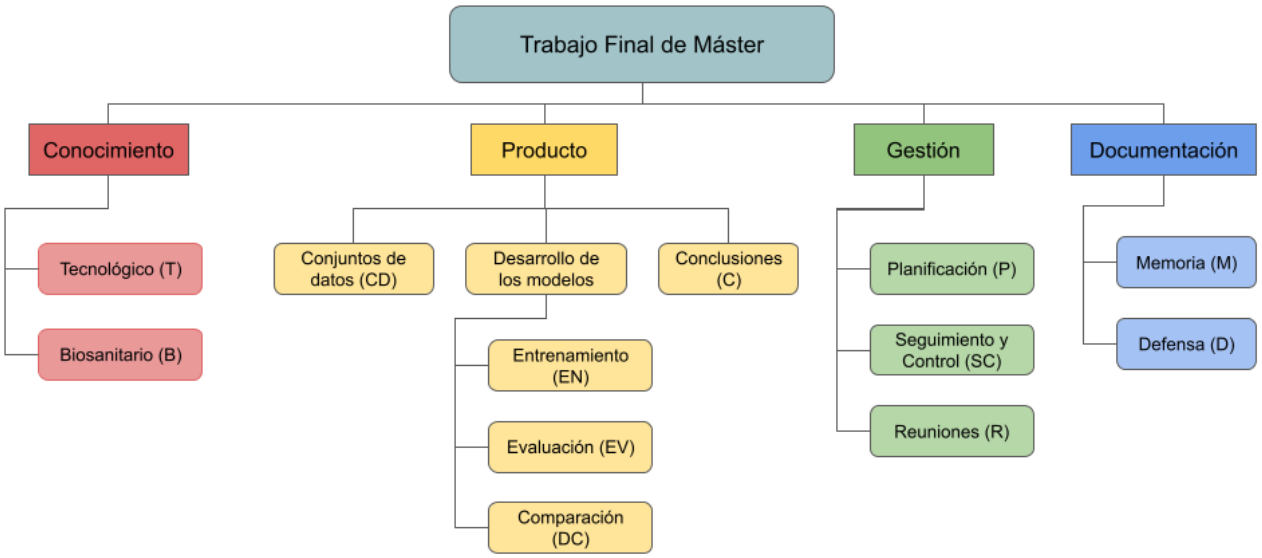


Figura 1.1: LDE diagrama

1.4.3. Gestión del tiempo

En este apartado se definen los límites para terminar el proyecto y entregar los entregables. En la imagen 1.2 podemos ver el diagrama de Gantt del proyecto, donde se estiman los tiempos de inicio y final de los paquetes de trabajo descritos previamente.

Figura 1.2: Diagrama de Gantt

Capítulo 2

Estado del Arte

2.1. Introducción

Este capítulo constituye una exploración del terreno científico relacionado con el uso del machine learning en el análisis de datos genéticos con el propósito de mejorar la prevención de la metástasis en pacientes con cáncer. Este análisis proporciona una visión de las investigaciones más relevantes y recientes en esta área.

El cáncer y su capacidad para propagarse y formar metástasis, representa uno de los mayores desafíos en la salud de la población mundial. En este contexto, la aplicación de técnicas de machine learning para el análisis de datos genéticos ofrece nuevas oportunidades para comprender mejor la propagación del cáncer, identificar biomarcadores de riesgo y desarrollar herramientas de detección temprana y predicción de metástasis precisas y efectivas.

En este capítulo, se examinarán los avances más significativos y las tendencias actuales en la aplicación del machine learning en el ámbito de la salud, especialmente relacionado con el cáncer y la metástasis. Este análisis permitirá fundamentar y contextualizar el desarrollo del presente Trabajo Final de Máster, proporcionando una base sólida para el desarrollo del proyecto y el cumplimiento de los objetivos planteados.

2.2. Justificación del Tema de Investigación

La metástasis es responsable de la mayoría de las muertes relacionadas con el cáncer [4] y su prevención efectiva es vital para mejorar la salud de los pacientes. En este contexto, aprovechando el desarrollo que se ha dado en la última década con la aparición de nuevas tecnologías cada vez más sofisticadas, como los algoritmos de machine learning, se han creado nuevas oportunidades para la aplicación de herramientas computacionales avanzadas en la investigación del ámbito sanitario.

De esta manera, diversos estudios han observado que estas herramientas son efectivas a la hora de diagnosticar, tratar e investigar enfermedades, partiendo de datos como la expresión génica [5]. Así mismo, en lo que al cáncer respecta, se han realizado diversas investigaciones que refuerzan la idea de que las herramientas de machine learning tienen mucho potencial para mejorar la identificación y los tratamientos de la enfermedad, como por ejemplo el uso de machine learning para la clasificación del tipo de cáncer usando la expresión génica [6, 7], la predicción de diferentes tipos de metástasis, como el cerebral [8] y el de hueso [9], la identificación de marcadores de metástasis en cáncer de mama [10] y la predicción de metástasis en cáncer de mama [11].

En todos los casos anteriormente mencionados, se ha empleado una variedad de algoritmos de aprendizaje automático, que van desde el Super Vector Classifier (SVC) hasta los modelos basados en redes neuronales, incluyendo también árboles de decisión. Sin embargo, investigaciones recientes [10, 11] han demostrado que los modelos desarrollados con redes neuronales exhiben una eficacia notablemente superior a la de otros modelos. Este hallazgo ha motivado la decisión de enfocar principalmente este Trabajo Final de Máster en el empleo de redes neuronales, en virtud de su rendimiento superior en comparación con otros enfoques de modelado.

Por otro lado, se han empleado datos de expresión génica como base para el desarrollo de los modelos. Estos datos han sido recopilados de diversas fuentes, entre las cuales se destaca el proyecto TCGA (The Cancer Genome Atlas) [12].

En conclusión, la aplicación del machine learning y el uso de datos génicos en la detección temprana y la predicción de la metástasis puede tener un impacto significativo en la salud pública al permitir intervenciones más efectivas y personalizadas.

2.3. Metodología

La estrategia escogida para este proyecto consiste en utilizar datos clínicos del proyecto TCGA (The Cancer Genome Atlas) [12] como principal fuente de información. TCGA es una iniciativa conjunta del Instituto Nacional del Cáncer (NCI) y del Instituto Nacional de Investigación del Genoma Humano (NHGRI) en los Estados Unidos, la cual tiene como objetivo mejorar nuestra comprensión del cáncer mediante la generación de un atlas comprensivo de las alteraciones genómicas que ocurren en una amplia variedad de tipos de cáncer.

El proyecto TCGA proporciona un extenso conjunto de datos clínicos de pacientes con cáncer, incluyendo datos de secuenciación del genoma, datos de expresión génica, y características clínicas de los pacientes, entre otros. Estos datos son recopilados y validados para asegurar su alta calidad y fiabilidad.

Al utilizar los datos del TCGA en este proyecto, se garantiza que los datos génicos dispo-

nibles para el desarrollo del trabajo son exhaustivos y diversos. Esto no solo proporciona una base sólida para la construcción de modelos de machine learning, sino que también permite una investigación más profunda y significativa en las características genómicas asociadas con la metástasis. Para garantizar la selección de datos de calidad, se llevará a cabo un análisis exhaustivo del conjunto de datos proporcionado por TCGA. Este proceso de análisis influirá directamente en la determinación del tipo específico de cáncer que será objeto de estudio en el Trabajo Final de Máster.

En cuanto a la metodología de desarrollo, se desarrollará el proyecto en cascada, siguiendo una secuencia de pasos posteriormente definida, avanzando a la siguiente fase una vez terminada la fase anterior. Aún así, en caso de ser necesario, se retrocederá a fases anteriores con la finalidad de mejorar la calidad del modelo de machine learning.

En la selección de la arquitectura adecuada para los modelos de machine learning en este proyecto, se prestará especial atención a las estructuras presentadas en estudios recientes. La justificación para el uso de diversas arquitecturas se basa en los resultados reportados en la literatura científica, que demuestran su eficacia en la clasificación y predicción de distintos tipos de cáncer utilizando datos de expresión génica.

Los **Support Vector Machines** (SVM) han mostrado un rendimiento notable en la clasificación de datos de expresión génica, alcanzando una precisión del 58 % en un estudio de clasificación de cáncer usando datos de expresión génica [6], y una precisión del 90 % en otro estudio de clasificación de cáncer de pulmón [13]. Este desempeño consistente sugiere que los SVM son eficaces para manejar la complejidad de los datos genómicos y pueden ser una herramienta robusta para este proyecto.

El modelo **Naive Bayes** (NGB) ha demostrado una alta precisión en estudios específicos. Por ejemplo, en la clasificación de adenocarcinoma y carcinoma de células del pulmón [13], alcanzó una precisión del 85 %. Además, en la predicción de metástasis cerebral [8], el área bajo la curva ROC fue de 0.845.

Los modelos **Random Forest** han mostrado una excelente capacidad de clasificación en estudios de cáncer. En la clasificación del tipo de cáncer de mama [14], se reportó una precisión del 90 %. Asimismo, en la predicción de metástasis cerebral [8], el área bajo la curva ROC fue de 0.858, superando a otros modelos. La robustez y precisión de los Random Forest los hacen ideales para tareas complejas de clasificación en este proyecto.

El modelo **XGBoost** también ha mostrado ser efectivo en la clasificación de datos de expresión génica. En el estudio "Cancer Classification of Gene Expression Data using Machine Learning Models" [6], XGBoost alcanzó una precisión del 64 %.

Las **Deep Neural Networks** (DNN) han demostrado un rendimiento excepcional en la predicción de cáncer. En el estudio sobre la predicción de cáncer de pulmón [15], los DNN

alcanzaron una precisión del 99 %. Para la predicción de metástasis ósea [9], los DNN lograron una precisión aproximada del 80 %. Además, en la predicción de metástasis cerebral [8], el área bajo la curva ROC fue de 0.839. Estos resultados sugieren que los DNN son altamente eficaces para capturar patrones complejos en los datos genómicos y proporcionar predicciones precisas.

La combinación de estas arquitecturas permite aprovechar sus fortalezas específicas para distintas tareas de clasificación y predicción en el proyecto. La implementación de SVM, Naïve Bayes, Random Forest, DNN y XGBoost no solo proporciona un marco sólido para el análisis de datos de expresión génica, sino que también maximiza la precisión y la robustez de los modelos desarrollados. Esta estrategia integral asegura que se utilicen los enfoques más efectivos y validados para el análisis de los datos clínicos proporcionados por TCGA, optimizando así los resultados del Trabajo Final de Máster.

2.4. Conclusiones

La metástasis, principal causa de mortalidad en pacientes con cáncer [1], representa un desafío crítico en el ámbito sanitario. En la última década, los avances en machine learning han promovido investigaciones efectivas en el diagnóstico y tratamiento del cáncer, especialmente al utilizar datos de expresión génica.

Los estudios indican el potencial de herramientas de machine learning en la identificación de biomarcadores y predicción de metástasis. Se destaca la superioridad observada en modelos basados en redes neuronales [10, 11].

Los datos de expresión génica, obtenidos de diversas fuentes, son esenciales en estas investigaciones, con el proyecto TCGA como una fuente principal [12]. Esta diversidad de datos permite una sólida base para la investigación.

En consecuencia, se enfocará en el análisis de datos clínicos del TCGA, influyendo en la elección del tipo de cáncer a estudiar. Se considerará la arquitectura de los modelos de machine learning que han mostrado mejoras significativas [6, 7, 8, 9, 10, 11, 13, 14, 15].

Capítulo 3

Tecnologías y métodos

En este apartado se explican las diferentes tecnologías utilizadas en el desarrollo del proyecto. Algunas de las tecnologías empleadas en el proyecto ya eran conocidas, lo que ha facilitado su uso. Por otro lado, también se han introducido nuevas tecnologías a lo largo del proyecto, lo que ha convertido su uso en un desafío.

3.1. Tecnologías usadas en el proyecto

3.1.1. R

R es un entorno de software libre para el cálculo estadístico y la creación de gráficos [16]. Se ha optado por utilizar este lenguaje debido a su capacidad para la recolección automática de datos del portal de TCGA, de donde se obtendrá el conjunto de datos con el que se trabajará en este proyecto.

3.1.1.1. Bioconductor y TCGAbiolinks

Bioconductor es un proyecto dedicado al desarrollo y difusión de software libre y código abierto que facilita el análisis de datos de ensayos biológicos [17]. Para este proyecto, es necesario utilizar Bioconductor para instalar los paquetes requeridos para la obtención y análisis de los conjuntos de datos proporcionados por TCGA.

TCGAbiolinks es un paquete que facilita la recuperación de datos de acceso abierto de TCGA, pudiendo preparar los datos utilizando las estrategias de preprocesamiento adecuadas, y proporcionando los medios para llevar a cabo diferentes tipos de análisis [18]. Este paquete ofrece diversas funciones útiles para obtener y filtrar los datos necesarios para construir un conjunto de datos con información génica de pacientes con cáncer de mama.

3.1.2. Python

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel muy conocido [19]. Se ha decidido realizar todo el desarrollo de aprendizaje automático de este Trabajo Final de Máster en el lenguaje de programación Python, debido a su facilidad de uso y a la amplia variedad de paquetes que ofrece. En los siguientes apartados se describen los principales paquetes utilizados en este trabajo.

3.1.2.1. Imbalanced-learn

Imbalanced-learn (importado en python como imblearn) es una biblioteca de código abierto con licencia MIT que proporciona herramientas para la clasificación de clases desequilibradas [20]. Esta biblioteca es necesaria debido a que el conjunto de datos obtenido está desbalanceado. Se profundiza más sobre este aspecto en el apartado 4.2.

3.1.2.2. Scikit-learn

Scikit-learn (importado como sklearn) es una biblioteca de código abierto que ofrece herramientas sencillas y eficaces para el análisis predictivo de datos [21]. Por un lado se utilizará este paquete para crear diferentes tipos de modelos como Naive Bayes y Regresión Logística, y por otro lado se usarán las herramientas de evaluación de modelos que ofrece. Además, también se utilizará para la búsqueda de hiperparámetros mediante GridSearchCV.

3.1.2.3. XGBoost

XGBoost es una librería de potenciación de gradiente optimizada para ser eficiente y flexible [22]. Esta biblioteca se utilizará para crear el modelo XGB.

3.1.2.4. Tensorflow

TensorFlow es una biblioteca de código abierto desarrollado por Google que ofrece diferentes utilidades para el desarrollo de modelos de aprendizaje automático [23]. En este trabajo se utilizará este paquete para desarrollar modelos DNN.

3.1.2.5. PyTorch

PyTorch es una biblioteca de código abierto que ofrece utilidades para aplicaciones de aprendizaje automático [24]. En este trabajo se utilizará esta biblioteca para usar el modelo TabNetClassifier.

3.1.2.6. AutoGluon

AutoGluon es una herramienta de código abierto desarrollada por Amazon que automatiza el proceso de entrenamiento y optimización de modelos de aprendizaje automático [25]. Es decir, además de seleccionar los hiperparámetros de manera automatica, también selecciona los modelos y realiza el ensamblaje de modelos.

3.1.2.7. Optuna

Optuna es un software de código abierto diseñado para la optimización automática de hiperparámetros [26]. Esta biblioteca será de utilidad en la búsqueda de hiperparámetros para los modelos XGB y DNN.

3.1.2.8. Bibliotecas de visualización de datos

Seaborn es una biblioteca de visualización de datos construida sobre Matplotlib que simplifica la creación de gráficos [27]. Se utilizarán estas dos biblioteca para visualizar la evaluación de los modelos.

3.2. Tecnologías para el desarrollo

3.2.1. RStudio

RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Es un entorno conveniente de usar gracias a diferentes características como la consola, el editor de sintaxis, las herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

3.2.2. Google Colab

Google Colaboratory es un servicio alojado de Jupyter Notebook que proporciona acceso gratuito a recursos informáticos, incluidas GPU y TPU, lo que es muy conveniente para el desarrollo de modelos de aprendizaje automático.

3.2.3. Google Drive

Google Drive es una plataforma para almacenar y compartir archivos en la nube. Por lo tanto, al ser un entorno conocido anteriormente, se ha utilizado para guardar copias de seguridad del proyecto, además de almacenar los conjuntos de datos y modelos generados a lo largo del

desarrollo. De esta manera, al ejecutar los programas en Google Colab, los conjuntos de datos y modelos son de fácil acceso.

3.3. Modelos Machine Learning

En esta sección se describen los modelos de machine learning que van a ser utilizados en el Trabajo Final de Máster.

3.3.1. Naive Bayes

El modelo Naive Bayes es un algoritmo de aprendizaje supervisado que se fundamenta en el teorema de Bayes. Este modelo asume de manera simplificada que existe independencia entre las características (predictores) en el conjunto de datos, es decir, que el valor de una característica no afecta al valor de otra. En este proyecto se trabaja con RNA-seq, donde el valor de un gen puede estar relacionado con el valor de otro gen, por lo tanto, se puede inferir que este modelo no tendrá un buen desempeño. Sin embargo, es interesante analizar su rendimiento para poder observar la mejora al implementar modelos más adecuados.

3.3.2. SVC

El modelo SVC (Support Vector Classifier) es un algoritmo de aprendizaje supervisado utilizado para la clasificación de datos el cual se basa en el concepto de máquinas de vectores de soporte (Support Vector Machines). El objetivo de este modelo es encontrar el hiperplano óptimo que mejor separe las diferentes clases en el espacio de características. Este hiperplano se elige de manera que maximice la distancia entre los puntos de datos más cercanos de las diferentes clases, lo que permite una mayor capacidad de generalización del modelo.

En el caso de conjuntos de datos no linealmente separables, el SVC proyecta los datos en un espacio dimensionalmente superior donde se vuelven linealmente separables. De esta manera es posible encontrar un hiperplano óptimo en ese espacio de características aumentado, resultando en una separación no lineal en el espacio de características original.

El SVC tiene parámetros importantes, como el parámetro de regularización (C) y el tipo de kernel, que pueden ajustarse para mejorar el rendimiento del modelo en diferentes conjuntos de datos. Además, es robusto frente a la presencia de datos atípicos en el conjunto de datos y puede manejar eficientemente conjuntos de datos de alta dimensionalidad, como es en el caso de los datos génicos.

3.3.3. Regresión Logística

La Regresión Logística es un modelo de aprendizaje automático utilizado principalmente para problemas de clasificación binaria, como la detección de metástasis. Aunque es más comúnmente aplicado a problemas binarios, este modelo puede extenderse para abordar problemas de clasificación multiclase. A diferencia de la regresión lineal, que se emplea para predecir valores numéricos, la regresión logística estima la probabilidad de que una observación pertenezca a una clase particular.

El modelo de Regresión Logística se basa en la función sigmoide, que transforma cualquier valor real en un rango de [0, 1]. Una vez entrenado, el modelo utiliza estas probabilidades predichas para tomar decisiones de clasificación. En el contexto de la detección de metástasis, esto significa predecir la probabilidad de que una muestra de RNA-seq indique la presencia de metástasis.

La Regresión Logística es un modelo flexible, fácil de interpretar y eficiente desde el punto de vista computacional, lo que la hace especialmente útil en la etapa inicial de un proyecto de detección de metástasis. Analizar su desempeño puede proporcionar una línea de base sólida, permitiendo comparaciones con modelos más complejos y adecuados que puedan implementarse posteriormente.

3.3.4. Random Forest

Random Forest es un algoritmo de aprendizaje supervisado que se utiliza tanto para problemas de regresión como de clasificación, el cual se basa en la construcción de múltiples árboles de decisión durante el entrenamiento y en la combinación de sus predicciones para obtener una predicción final más robusta y precisa. Cada árbol se genera utilizando un subconjunto aleatorio de las características del conjunto de datos y un subconjunto aleatorio de las observaciones de entrenamiento mediante muestreo con reemplazo. Este proceso introduce diversidad en los árboles individuales y reduce el riesgo de sobreajuste. Para problemas de clasificación, como la detección de metástasis, la predicción final se determina por mayoría de votos entre todos los árboles del bosque.

Debido a su naturaleza de conjunto y a la aleatoriedad en la construcción de árboles, Random Forest tiende a ser robusto contra el sobreajuste y generaliza bien a datos no vistos. En el caso de la detección de metástasis a partir de RNA-seq, este algoritmo puede proporcionar un modelo robusto y preciso, mejorando la capacidad de detectar patrones complejos en los datos genómicos. La comparación del rendimiento de Random Forest con otros modelos puede ofrecer valiosas perspectivas sobre la mejora de la precisión en la detección de metástasis.

3.3.5. XGBoost

XGBoost es una biblioteca de código abierto ampliamente utilizada para problemas de aprendizaje supervisado como clasificación. Utiliza árboles de decisión más complejos que otros métodos de boosting, lo que permite capturar mejor las relaciones intrincadas en los datos. El algoritmo implementa técnicas de regularización para evitar el sobreajuste, asegurando que el modelo generalice bien a datos nuevos. Además, XGBoost construye una secuencia de árboles de decisión, cada uno corrigiendo los errores de los anteriores, mejorando continuamente el rendimiento del modelo [28].

3.3.6. Deep Neural Network

Una Red Neuronal Profunda es un tipo avanzado de modelo de aprendizaje automático inspirado en la estructura y función del cerebro humano. En el contexto de este proyecto, las DNNs pueden ser extremadamente útiles debido a su capacidad para aprender representaciones complejas y abstractas de los datos genómicos.

La estructura de una DNN comienza con la capa de entrada, que recibe directamente los datos, donde cada nodo representa una característica de estos datos. A continuación, se encuentran las capas ocultas, que pueden ser una o más, cada una compuesta por nodos interconectados. Estas capas ocultas realizan la mayor parte del procesamiento y abstracción de la información. Finalmente, la capa de salida produce la predicción final, cuyo tamaño depende del problema en cuestión, como la clasificación binaria para detectar la presencia de metástasis.

Durante el proceso de entrenamiento, los pesos que conectan los nodos se inicializan con valores pequeños y aleatorios. Estos pesos se actualizan iterativamente a lo largo de las épocas hasta que la red converge a una solución óptima o alcanza un criterio de parada predefinido (early stopping). Aunque este tipo de modelo es capaz de aprender relaciones complejas, también presenta desventajas, como la necesidad de grandes conjuntos de datos para un entrenamiento efectivo. Esto supone un reto, ya que el conjunto de datos utilizado en este trabajo no es muy grande.

3.3.7. TabNet

TabNet es un modelo de aprendizaje profundo diseñado para datos tabulares que utiliza un mecanismo de atención para seleccionar y procesar las características más relevantes de los datos, capturando relaciones complejas y mejorando la precisión predictiva. Esta perspectiva es útil, ya que los datos RNA-seq se procesan como datos tabulares.

A diferencia de las redes neuronales tradicionales, TabNet maneja eficientemente la estructura de los datos tabulares y previene el sobreajuste mediante técnicas de regularización.

Además, su capacidad de interpretación permite entender qué características influyen más en las predicciones, lo cual puede ayudar a identificar genes clave en la detección de metástasis, proporcionando información valiosa para futuras investigaciones.

3.3.8. AutoGluon: TabularPredictor

TabularPredictor de AutoGluon es una herramienta automatizada de aprendizaje automático diseñada para trabajar con datos tabulares. Utiliza una variedad de algoritmos y técnicas avanzadas de machine learning, incluyendo ensamblaje de modelos, para optimizar automáticamente el rendimiento predictivo. Al igual que en el caso anterior, esta herramienta es relevante ya que los datos RNA-seq se procesan como datos tabulares. Por otro lado, también ofrece la oportunidad de identificar las características claves en la detección de metástasis.

3.4. Métricas y herramientas de evaluación

Para medir la efectividad de los modelos de aprendizaje automático que se van a desarrollar, es crucial emplear métricas de evaluación precisas y adecuadas. Estas métricas no solo permiten evaluar y comparar diferentes modelos, sino que también proporcionan una base sólida para el análisis cuantitativo que guía la elección del modelo final.

3.4.1. AUC-ROC

La **AUC-ROC** es fundamental para evaluar la capacidad de un modelo de clasificación binaria. Esta métrica considera tanto la tasa de verdaderos positivos como la tasa de falsos positivos, proporcionando una visión integral de la discriminación del modelo. Un valor de AUC cercano a 1 indica un excelente desempeño en distinguir entre las clases de metástasis y no metástasis.

3.4.2. Matriz de confusión

La **matriz de confusión** es otra herramienta esencial en la evaluación de modelos de clasificación. Esta matriz permite visualizar el desempeño del modelo mostrando las verdaderas predicciones positivas y negativas junto con las falsas predicciones positivas y negativas. Es especialmente útil para entender cuántos casos de metástasis se detectaron correctamente y cuántos se perdieron.

3.4.3. Precisión, Recall y F1-score

Por último están las métricas de **precisión**, **recall** y **F1-score**. La precisión es la proporción de verdaderos positivos sobre el total de predicciones positivas hechas por el modelo. El recall, por otro lado, mide los verdaderos positivos sobre el total de casos positivos, es decir, la capacidad del modelo para identificar todos los casos de metástasis. El F1-score es la media armónica de la precisión y el recall, y ofrece un balance entre ambas métricas, siendo especialmente útil cuando se desea equilibrar la precisión y el recall en situaciones donde las clases están desbalanceadas. En este trabajo, es fundamental lograr altos valores de recall y F1-score para la clase de metástasis, dado que el objetivo principal es la detección precisa de metástasis para garantizar el tratamiento adecuado del paciente.

Capítulo 4

Procesamiento de datos

4.1. Sobre el dataset

El objetivo de este proyecto es predecir la metástasis en casos de cáncer utilizando datos de expresión génica. Para ello, se recopilarán y analizarán los datos genómicos disponibles en el portal de datos del The Cancer Genome Atlas (TCGA). Esta base de datos proporciona información detallada sobre la expresión de miles de genes en muestras tumorales y metastáticos.

Para seleccionar el tipo de cáncer a estudiar, se deben considerar factores como la disponibilidad de datos y el balance de los mismos. En la siguiente tabla se muestra la cantidad de casos disponibles de metástasis por tipo de cáncer con datos de RNA-seq:

Tipo de cáncer	Casos de tumor metastático	Casos totales
Cáncer de piel	398	499
Cáncer de próstata	102	639
Cáncer de tiroides	56	945
Cáncer de mama	45	1383
Cáncer de riñón	35	1408
Cáncer de páncreas	15	138
Cáncer de pulmón	12	246
Cáncer de colon	11	270
Cáncer de esófago	11	50

Tabla 4.1: Casos disponibles de metástasis por tipo de cáncer

Al observar la tabla, se puede notar que una gran cantidad de casos de cáncer de piel son metastásicos, lo que puede implicar que identificar la metástasis en este tipo de cáncer no represente un gran desafío. Considerando estos datos y mi interés particular en el cáncer de mama, se ha decidido enfocar el estudio en dos tipos de cáncer con un número significativo de casos de metástasis: el cáncer de próstata y el cáncer de mama.

Para el desarrollo de este Trabajo de Fin de Máster, se utilizarán casos etiquetados como "tumor primario" y "tumor metastático". Después de filtrar los casos especiales y los no etiquetados, las cantidades de cada caso disponibles para el estudio se detallan en las tablas 4.2 y 4.3, así como el proyecto de origen. Los datos de cáncer de próstata se han obtenido de los siguientes tres proyectos:

1. **Prostate Adenocarcinoma (TCGA-PRAD)**

2. **Count Me In: The Metastatic Prostate Cancer (CMI-MPC)**

3. **Genomic Characterization of Metastatic Castration Resistant Prostate Cancer (WCDT-MCRPC)**

Proyecto	Casos de tumor primario	Casos de tumor metastático	Total
TCGA-PRAD	501	1	502
CMI-MPC	38	0	38
WCDT-MCRPC	0	99	99
Total:	539	100	639

Tabla 4.2: Cantidad de casos de cáncer de próstata

En el caso del cáncer de mama, los datos se han obtenido de los siguientes tres proyectos:

1. **Breast Invasive Carcinoma (TCGA-BRCA)**

2. **The Metastatic Breast Cancer Project (CMI-MBC)**

3. **Count Me In: The Angiosarcoma Project (CMI-ASC)**

Proyecto	Casos de tumor primario	Casos de tumor metastático	Total
TCGA-BRCA	1111	7	1118
CMI-MBC	0	32	32
CMI-ASC	0	1	1
Total:	1111	40	1151

Tabla 4.3: Cantidad de casos de cáncer de mama

4.2. Construcción del dataset

Para la creación del dataset de este Trabajo de Fin de Máster, se ha utilizado el paquete TCGAbiolinks de R para descargar los datos de expresión génica desde el portal de datos del The Cancer Genome Atlas (TCGA). Los datos de RNA-seq disponibles en este portal se presentan en varias métricas, las cuales son utilizadas para cuantificar y comparar la expresión génica entre diferentes tipos de muestra en estudios de transcriptómica:

- **Unstranded**: Esta medida de la expresión génica no tiene en cuenta la dirección de las lecturas obtenidas durante el proceso de secuenciación. Esto significa que no se diferencia entre lecturas que se originaron en el sentido positivo o negativo de la cadena de ADN.
- **Stranded First**: Esta métrica de expresión génica tiene en cuenta la dirección de la primera lectura. Es decir, distingue entre lecturas que se originaron en el sentido positivo o negativo de la cadena de ADN, basándose en la dirección de la primera lectura.
- **Stranded Second**: Esta medida considera la dirección de la segunda lectura obtenida durante el proceso de secuenciación.
- **TPM (Transcripts Per Million)**: Normaliza la cantidad de transcritos por millón de lecturas. Esta normalización permite comparaciones directas entre muestras, ya que tiene en cuenta la profundidad de secuenciación y el tamaño del gen.
- **FPKM (Fragments Per Kilobase of transcript per Million mapped reads)**: Normaliza la cantidad de fragmentos por kilobase de transcrito por millón de lecturas mapeadas. Esta métrica tiene en cuenta la longitud del transcrito y la profundidad de secuenciación para normalizar los datos de expresión génica.
- **FPKM-UQ (Upper Quartile normalization)**: Esta medida es una variante del FPKM que usa la normalización del cuartil superior, con la finalidad de reducir el impacto de los valores extremos, mejorando así la estabilidad y la fiabilidad de las comparaciones entre muestras.

De estas métricas, se ha optado por utilizar la métrica TPM (Transcripts Per Million). Esta elección se basa en su capacidad para normalizar los datos de expresión génica de manera efectiva, permitiendo comparaciones directas entre muestras y corrigiendo las diferencias en la profundidad de secuenciación y el tamaño del gen.

De este modo, se ha creado un archivo CSV para cada tipo de cáncer y cada proyecto, donde cada fila representa un caso y cada columna representa un gen. Finalmente, se crean dos archivos CSV por tipo de cáncer, uno con los casos de tumor primario y otro con los casos de tumor metastático. En la tabla 4.4 se pueden observar las dimensiones de cada archivo. Partiendo de estos CSV, como se explica en el apartado 4.3, se hará la limpieza de datos y se crearán los conjuntos de datos necesarios para la creación de los modelos predictivos. El proceso completo de recopilación de datos y creación de archivos a través de R se detalla en el código del anexo A.1.

Archivo	Filas	Columnas	Elementos
PRAD-TP_genes.csv	539	60.661	32.696.279
PRAD-TM_genes.csv	100	60.661	6.066.100
BRCA-TP_genes.csv	1.111	60.661	67.394.371
BRCA-TM_genes.csv	40	60.661	2.426.440

Tabla 4.4: Dimensiones de los archivos

4.3. Conjunto de datos

A partir de los archivos CSV combinados, se ha procedido a la limpieza de datos, eliminando genes con baja expresión y normalizando los datos mediante una transformación logarítmica. Este paso es crucial para asegurar la calidad y comparabilidad de los datos en los modelos predictivos. La detección de genes con baja expresión se ha realizado de forma separada en los conjuntos de datos de tumor primario y tumor metastático, descartando únicamente aquellos genes con baja expresión en ambos conjuntos. Esta metodología se debe a la desbalanceada cantidad de datos, que podría llevar a la eliminación de genes significativos para los casos de metástasis si se analizaran de manera conjunta, evitando así el descarte erróneo de genes relevantes. De esta manera, se ha conseguido disminuir significativamente la cantidad de columnas, como se observa en la siguiente tabla:

Tipo de cáncer	Filas	Columnas	Elementos
Cáncer de próstata	639	36.735	23.473.665
Cáncer de mama	1.151	27.606	31.774.506

Tabla 4.5: Dimensiones de los datos

El objetivo de este proyecto es crear modelos de predicción para cada tipo de cáncer, por lo que los conjuntos de datos estarán separados por tipo de cáncer. Esta separación permi-

te desarrollar modelos específicos y más precisos para cada tipo de cáncer, aprovechando las características particulares de la expresión génica en los distintos tipos de tumor.

Una vez completada la limpieza y normalización de los datos, se ha procedido a crear los conjuntos de entrenamiento, validación y prueba. El conjunto de prueba constituye el 25 % del total de los datos, mientras que el conjunto de validación corresponde al 25 % del 75 % restante, es decir, al 18.75 % del total de los datos. En la siguiente tabla podemos observar las proporciones de los tipos de tumor para cada tipo de cáncer:

Tipo de cáncer	Proporción de TP	Proporción de TM
Cáncer de próstata	84.4 %	15.6 %
Cáncer de mama	96.5 %	3.5 %

Tabla 4.6: Proporciones de las clases para cada tipo de cáncer

Dado que los datos están significativamente desbalanceados, se han aplicado diferentes técnicas para manejar el desbalanceo en cada conjunto:

- **Conjunto de entrenamiento**: Para este conjunto, se generarán datos sintéticos utilizando tres métodos distintos: ADASYN (Adaptive Synthetic Sampling), SMOTETomek (una combinación de SMOTE y Tomek links), y GAN (Generative Adversarial Networks). Estas técnicas permiten aumentar la cantidad de casos en la clase minoritaria (tumor metastático) para equilibrar el conjunto de datos, mejorando así la capacidad del modelo para aprender a detectar ambas clases de manera efectiva.

- **Conjunto de validación**: Se seleccionarán aleatoriamente casos de la clase mayoritaria (tumor primario) para igualar la cantidad de casos de la clase minoritaria (tumor metastático). Esta estrategia garantiza que durante la validación, ambas clases tengan el mismo peso, evitando que el modelo favorezca la detección de la clase mayoritaria. La métrica de mayor interés en este contexto es el recall de la clase metástasis, ya que el objetivo principal es detectar correctamente los casos de metástasis.

- **Conjunto de prueba**: Este conjunto se mantendrá sin modificaciones para reflejar la distribución real de los datos. Esto es importante para evaluar el desempeño del modelo en condiciones reales, proporcionando una estimación precisa de su capacidad predictiva en un entorno realista.

Esta metodología asegura que el modelo sea robusto y equitativo en la detección de ambas clases, con un enfoque particular en la identificación de los casos de metástasis, lo cual es crucial para el objetivo final del proyecto.

4.3.1. Conjunto de validación

Una vez creados los tres conjuntos de datos para cada tipo de cáncer, se emplea una función específica para balancear los datos del conjunto de validación (4.1). Esta función toma como parámetros el conjunto de datos y sus etiquetas, y devuelve estos mismos valores filtrados y balanceados en función de la clase minoritaria.

```
def balance_classes(X, y):
    class_0_indices = y[y == 0].index
    class_1_indices = y[y == 1].index
    n_samples = min(len(class_0_indices), len(class_1_indices))
    balanced_indices = np.concatenate([
        np.random.choice(class_0_indices, n_samples, replace=False),
        np.random.choice(class_1_indices, n_samples, replace=False)
    ]).tolist()
    # Filtrar las filas de X e y utilizando indices seleccionados
    X_balanced = X.loc[balanced_indices]
    y_balanced = y.loc[balanced_indices]
    return X_balanced, y_balanced
```

Programa 4.1: Función para balancear el conjunto de validación

En este punto, las proporciones y la cantidad de casos por cada conjunto para cada tipo de cáncer son las siguientes:

Tipo de cáncer	Conjunto	Casos de TP	Casos de TM
Cáncer de próstata	Train	303	56
	Validation	19	19
	Test	135	25
Cáncer de mama	Train	625	22
	Validation	8	8
	Test	278	10

Tabla 4.7: Cantidad de casos por conjunto

4.3.2. Conjunto de entrenamiento

Para el conjunto de entrenamiento, se van a crear tres tipos de conjuntos diferentes. Los dos primeros conjuntos se han creado usando los paquetes ADASYN y SMOTETomek (código 4.2).

```
smote_tomek = SMOTETomek(random_state=random_state)
X_train_st, y_train_st = smote_tomek.fit_resample(X_train, y_train)
adasyn = ADASYN(random_state=random_state)
X_train_ada, y_train_ada = adasyn.fit_resample(X_train, y_train)
```

Programa 4.2: Función para balancear el conjunto de validación

Para crear el tercer conjunto, se implementan dos redes neuronales: una generativa y una discriminativa, formando así una Red Generativa Adversarial (GAN). La red generativa, conocida como el generador, tiene la tarea de producir datos sintéticos a partir de un vector de ruido aleatorio. Por otro lado, la red discriminativa, o discriminador, se encarga de evaluar estos datos generados, distinguiendo entre los datos sintéticos y los datos reales de metástasis.

Durante el proceso de entrenamiento, las dos redes neuronales se enfrentan en un juego donde la mejora de una implica la mejora de la otra, manteniendo así un equilibrio. El generador se ajusta para crear datos cada vez más realistas, mientras que el discriminador se perfecciona para identificar con mayor precisión los datos falsos. El objetivo final es que el generador llegue a producir datos tan similares a los reales que el discriminador no pueda distinguir entre los datos auténticos y los generados. El proceso de creación de las redes y su entrenamiento se detalla en el código del anexo A.3. La siguiente tabla recoge las cantidades finales de casos por método utilizado para su creación.

Tipo de cáncer	Método	Casos de TP	Casos de TM
Cáncer de próstata	ADASYN	303	303
	SMOTETomek	303	303
	GAN	303	303
Cáncer de mama	ADASYN	625	626
	SMOTETomek	625	625
	GAN	625	625

Tabla 4.8: Cantidad de casos por método

Capítulo 5

Resultados

En este capítulo, se profundiza en los procedimientos de entrenamiento y evaluación de cada modelo desarrollado. Cada modelo se adapta específicamente a los conjuntos de datos creados con diferentes técnicas de balanceo. Esta estrategia nos permite explorar cómo cada tipo de modelo responde a conjuntos de datos sintéticos generados de manera diferente, proporcionando una visión más completa del rendimiento de los modelos en situaciones de desbalanceo de clases.

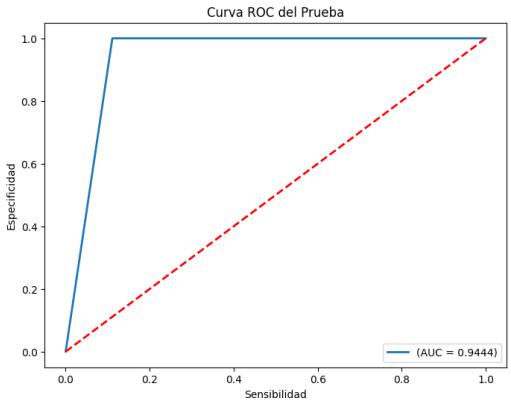
Para cada modelo, se describen los datos utilizados durante el entrenamiento, así como los hiperparámetros seleccionados para su configuración. A lo largo de esta sección, se presentan los resultados detallados obtenidos en la evaluación del conjunto de prueba. Estos resultados se analizan mediante métricas estándar, incluyendo la exactitud (accuracy), la pérdida (loss), la matriz de confusión y la curva ROC, lo que proporciona una comprensión completa del rendimiento de cada modelo y su capacidad para generalizar a nuevos datos de pacientes con cáncer.

5.1. Naive Bayes

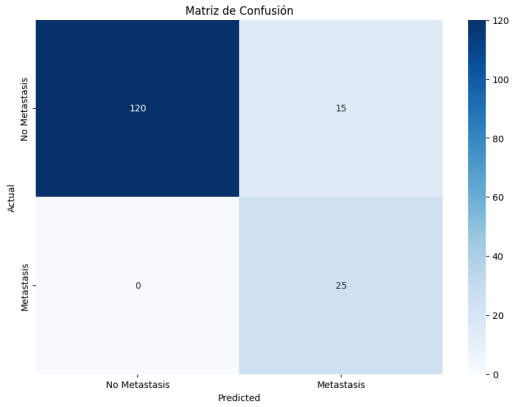
Se ha utilizado el modelo Gaussian Naive Bayes sin necesidad de ajustar hiperparámetros, ya que este modelo no requiere la misma sintonización extensiva que los próximos modelos. En este caso, se han juntado los conjuntos de entrenamiento y validación, utilizando el conjunto de prueba para la evaluación del modelo.

5.1.1. Evaluación del modelo

5.1.1.1. Conjunto de datos SMOTETomek



(a) Curva ROC

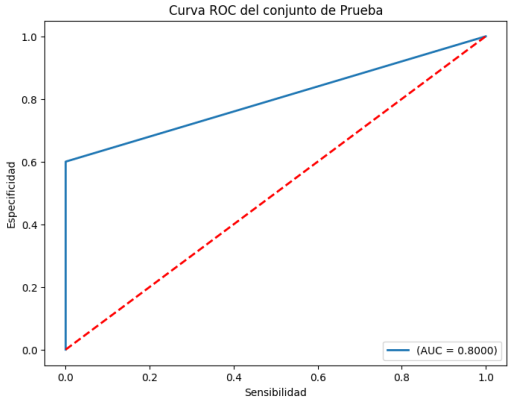


(b) Matriz de confusión

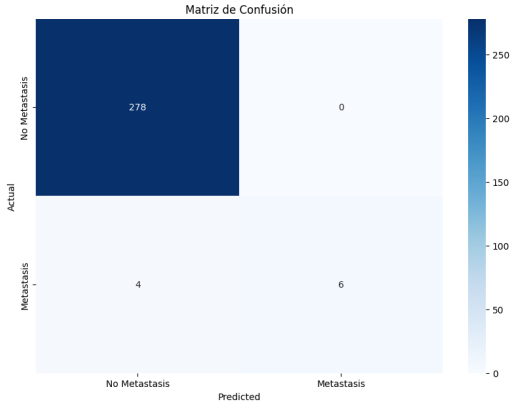
Figura 5.1: Resultados del modelo NB-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.89	0.94
Metástasis	0.62	1.00	0.77

Tabla 5.1: Métricas del modelo NB-ST para cáncer de próstata



(a) Curva ROC



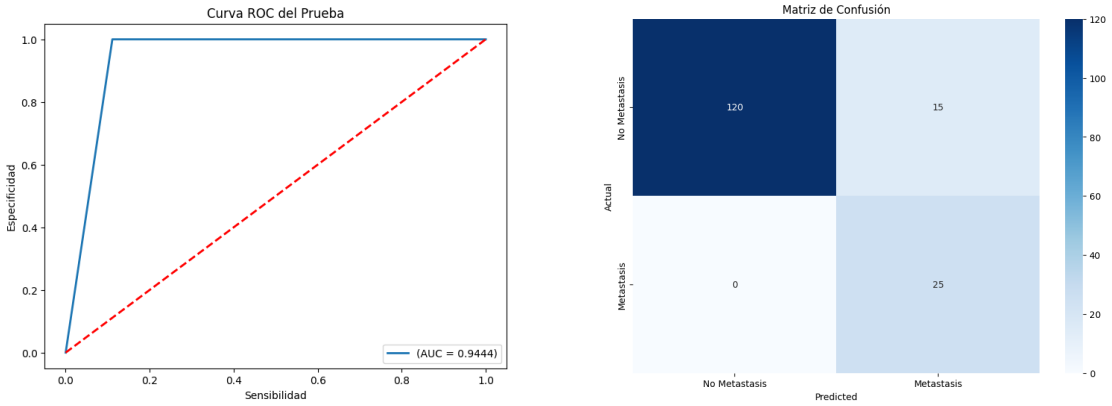
(b) Matriz de confusión

Figura 5.2: Resultados del modelo NB-ST para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	0.99	1.00	0.99
Metástasis	1.00	0.60	0.75

Tabla 5.2: Métricas del modelo NB-ST para cáncer de mama

5.1.1.2. Conjunto de datos ADASYN

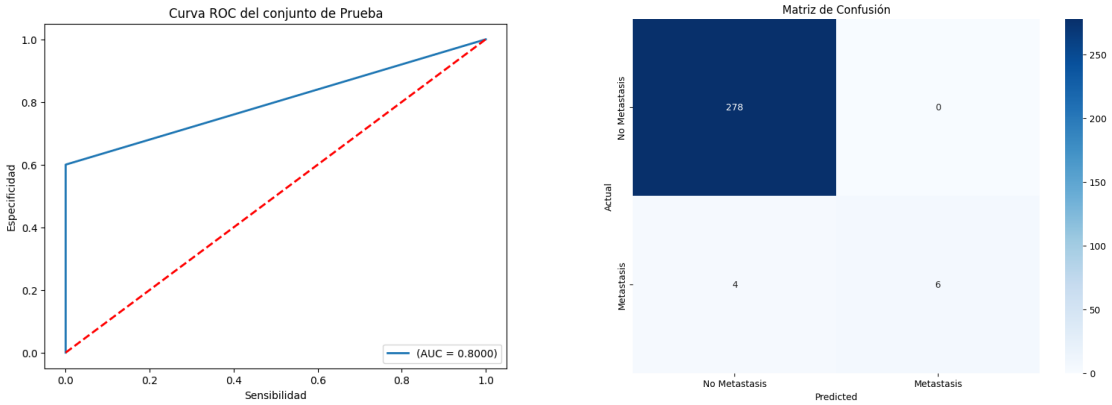


(a) Curva ROC (b) Matriz de confusión

Figura 5.3: Resultados del modelo NB-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.89	0.94
Metástasis	0.62	1.00	0.77

Tabla 5.3: Métricas del modelo NB-ADA para cáncer de próstata



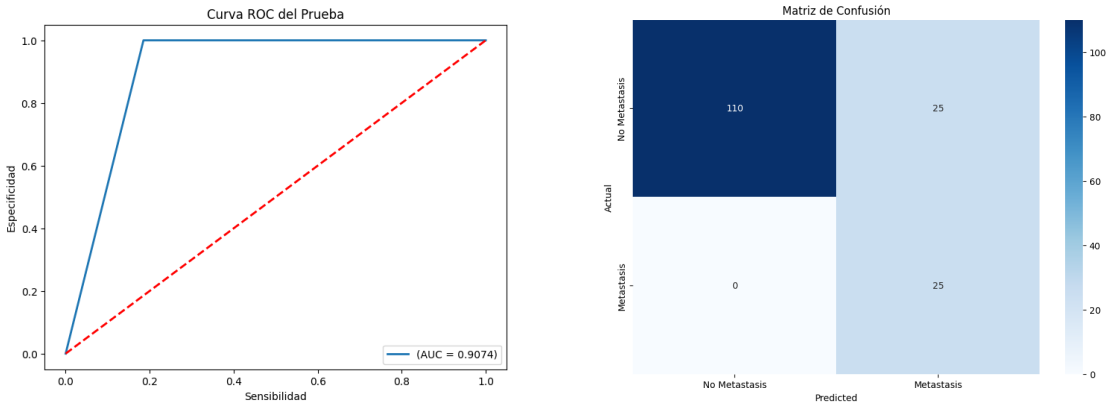
(a) Curva ROC (b) Matriz de confusión

Figura 5.4: Resultados del modelo NB-ADA para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	0.99	1.00	0.99
Metástasis	1.00	0.60	0.75

Tabla 5.4: Métricas del modelo NB-ADA para cáncer de mama

5.1.1.3. Conjunto de datos GAN

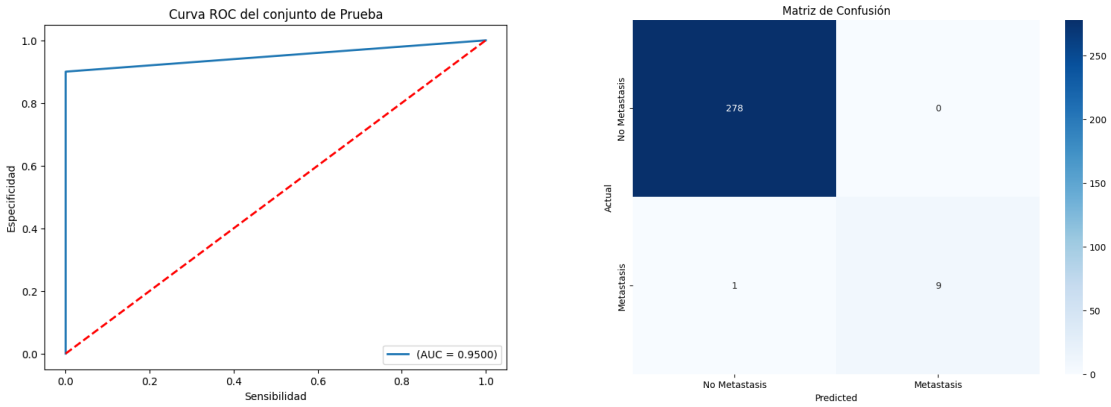


(a) Curva ROC (b) Matriz de confusión

Figura 5.5: Resultados del modelo NB-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.81	0.90
Metástasis	0.50	1.00	0.67

Tabla 5.5: Métricas del modelo NB-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.6: Resultados del modelo NB-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.6: Métricas del modelo NB-GAN para cáncer de mama

5.2. SVC

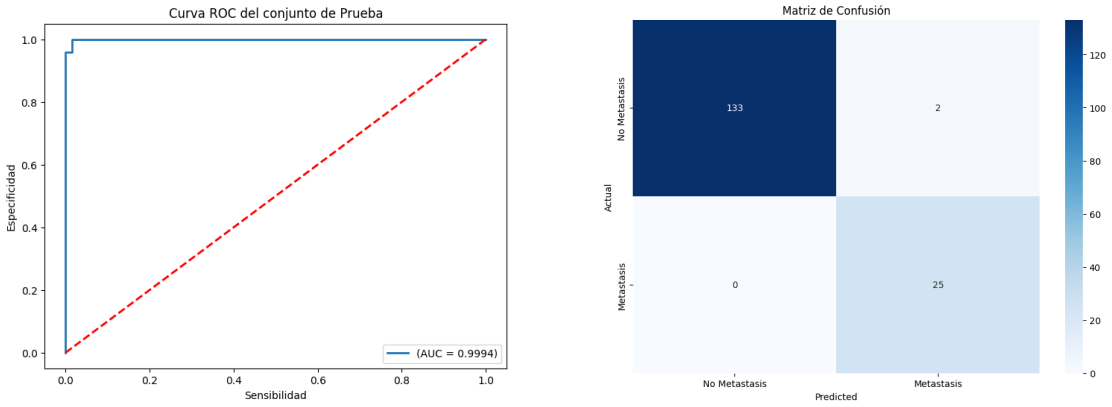
En el contexto del modelo de Support Vector Classifier (SVC), los hiperparámetros utilizados en el proceso de entrenamiento son los siguientes:

- **C**: Los valores utilizados son 0.01, 0.1, 1, 10 y 100.
- **Kernel**: define la función de transformación utilizada. En este caso se le ha dado los valores de "linear", "poly" y "rbf".
- **Gamma**: Los valores utilizados son "scale" y "auto".

Por otro lado, en este caso se han juntado los conjuntos de entrenamiento y validación, y se ha realizado validación cruzada con KFold. Además, se ha usado la métrica f1 como métrica de evaluación. Los mejores resultados se han obtenido

5.2.1. Evaluación del modelo

5.2.1.1. Conjunto de datos SMOTETomek

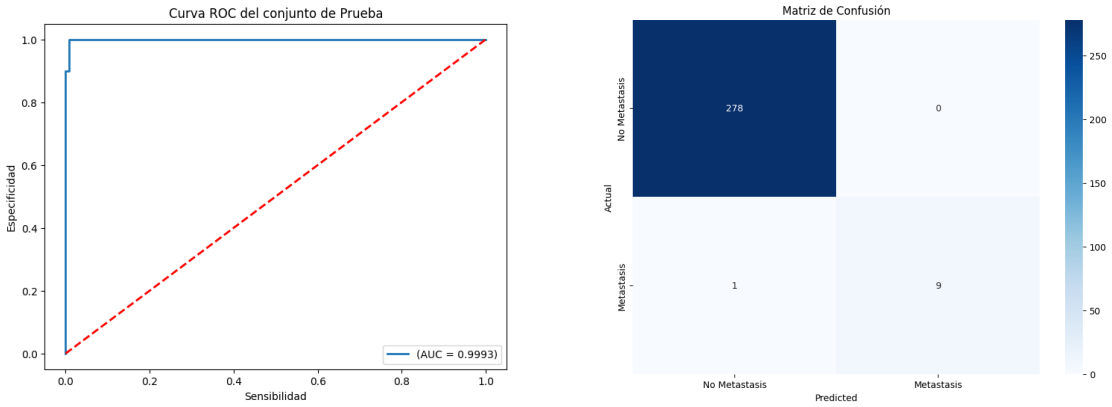


(a) Curva ROC (b) Matriz de confusión

Figura 5.7: Resultados del modelo SVC-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.99	0.99
Metástasis	0.93	1.00	0.96

Tabla 5.7: Métricas del modelo SVC-ST para cáncer de próstata



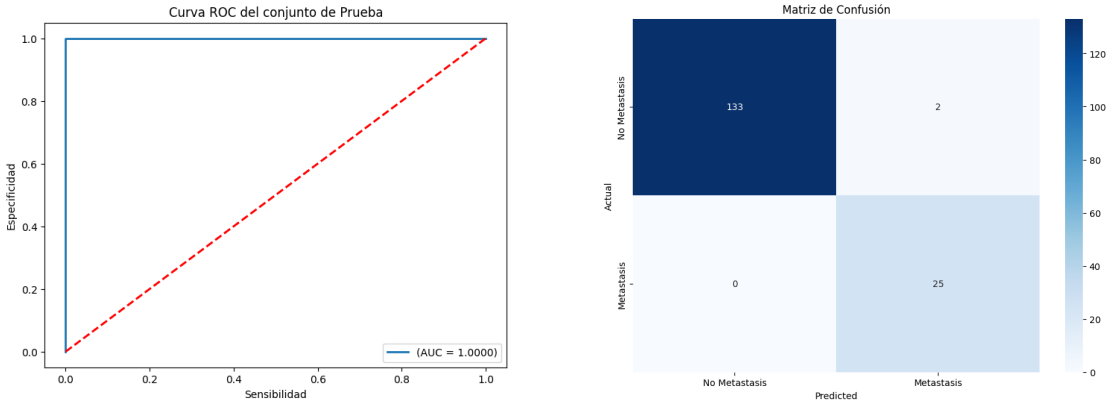
(a) Curva ROC (b) Matriz de confusión

Figura 5.8: Resultados del modelo SVC-ST para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.8: Métricas del modelo SVC-ST para cáncer de mama

5.2.1.2. Conjunto de datos ADASYN

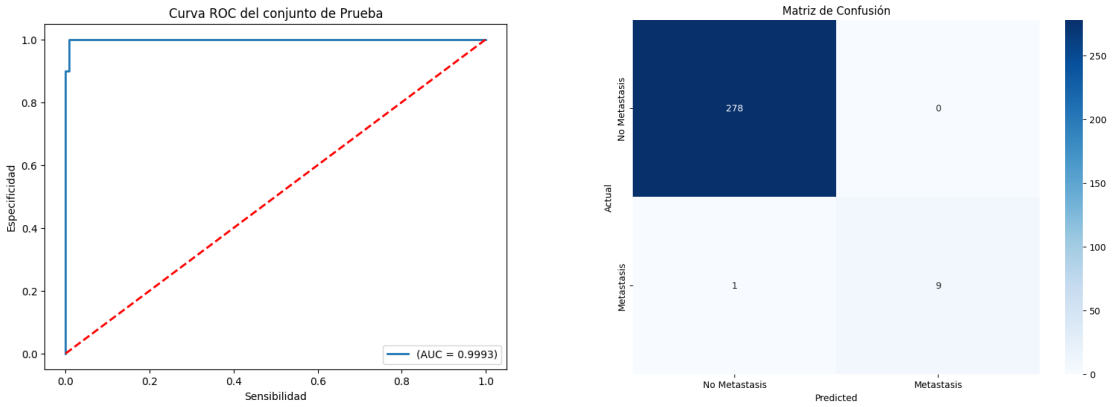


(a) Curva ROC (b) Matriz de confusión

Figura 5.9: Resultados del modelo SVC-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.99	0.99
Metástasis	0.93	1.00	0.96

Tabla 5.9: Métricas del modelo SVC-ADA para cáncer de próstata



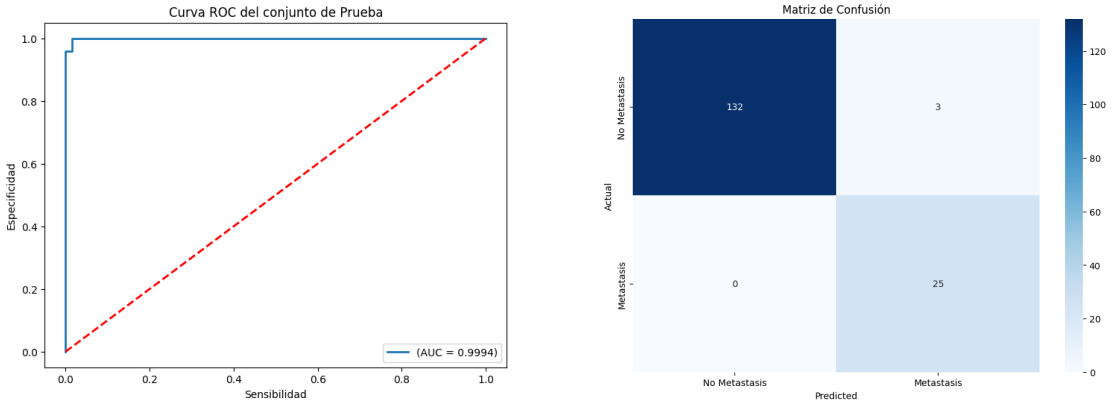
(a) Curva ROC (b) Matriz de confusión

Figura 5.10: Resultados del modelo SVC-ADA para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.10: Métricas del modelo SVC-ADA para cáncer de mama

5.2.1.3. Conjunto de datos GAN

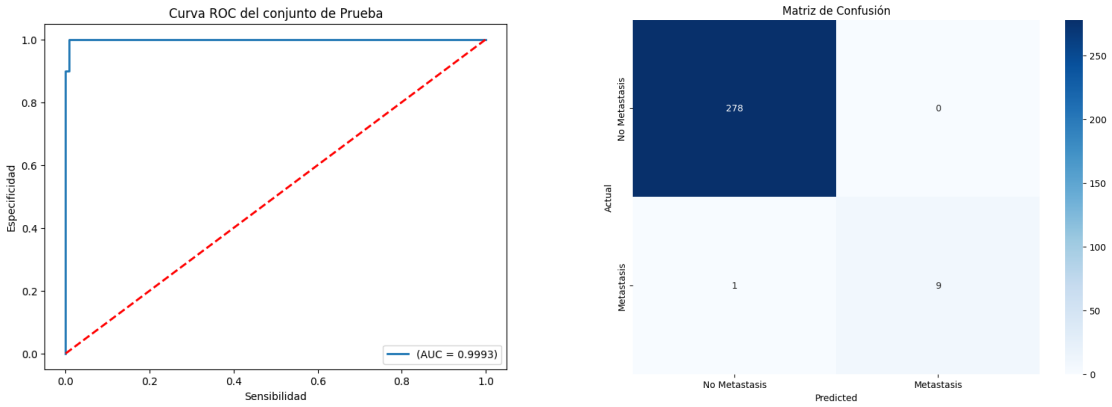


(a) Curva ROC (b) Matriz de confusión

Figura 5.11: Resultados del modelo SVC-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.11: Métricas del modelo SVC-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.12: Resultados del modelo SVC-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.12: Métricas del modelo SVC-GAN para cáncer de mama

5.3. Regresión logística

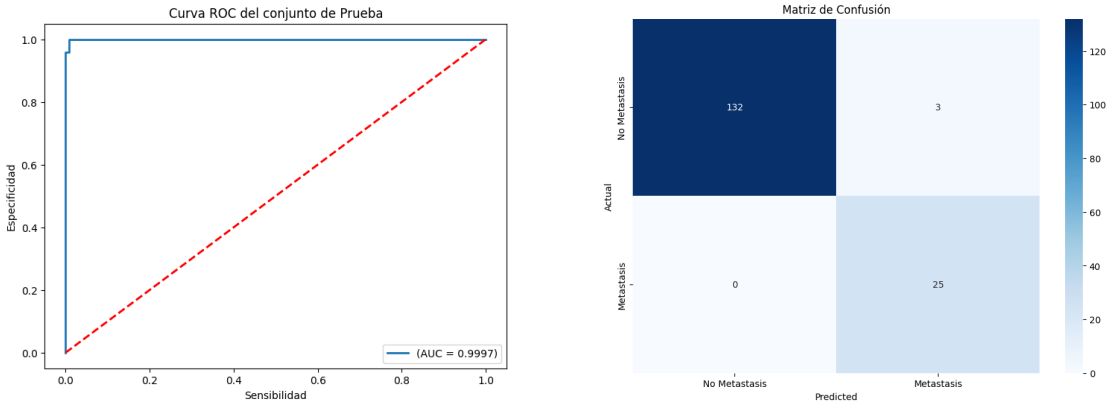
En el contexto del modelo de Regresión Logística (Logistic Regression), los hiperparámetros utilizados en el proceso de entrenamiento son los siguientes:

- **max_iter**: Los valores utilizados son 100, 500 y 1000. Este hiperparámetro controla el número máximo de iteraciones para el algoritmo de optimización.
- **C**: Los valores utilizados son 0.01, 0.1, 1, 10 y 100. Este hiperparámetro es el inverso de la regularización y controla la fuerza de regularización aplicada al modelo.

Por otro lado, en este caso se han juntado los conjuntos de entrenamiento y validación, y se ha realizado validación cruzada con KFold.

5.3.1. Evaluación del modelo

5.3.1.1. Conjunto de datos SMOTETomek

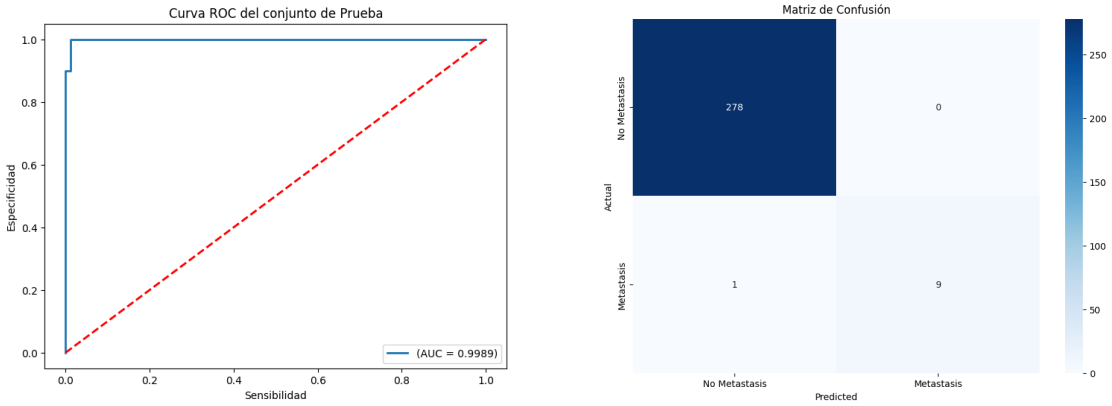


(a) Curva ROC (b) Matriz de confusión

Figura 5.13: Resultados del modelo LR-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.13: Métricas del modelo LR-ST para cáncer de próstata



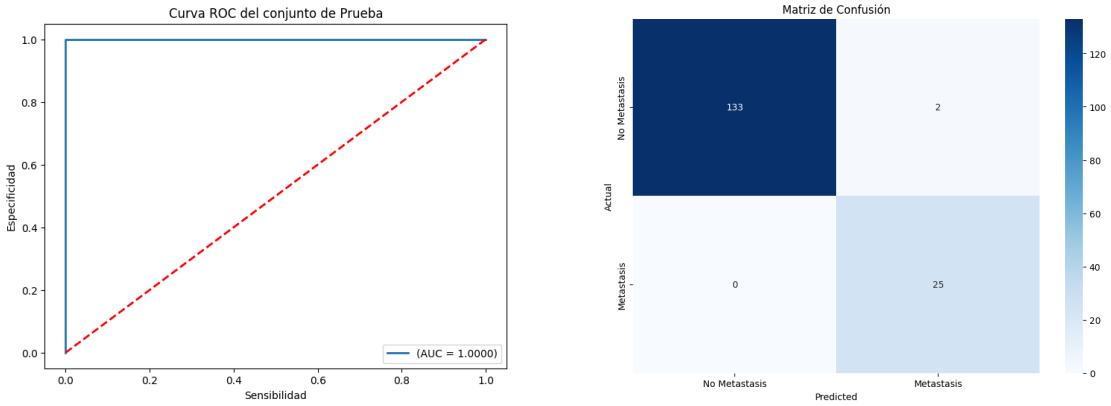
(a) Curva ROC (b) Matriz de confusión

Figura 5.14: Resultados del modelo LR-ST para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.14: Métricas del modelo LR-ST para cáncer de mama

5.3.1.2. Conjunto de datos ADASYN

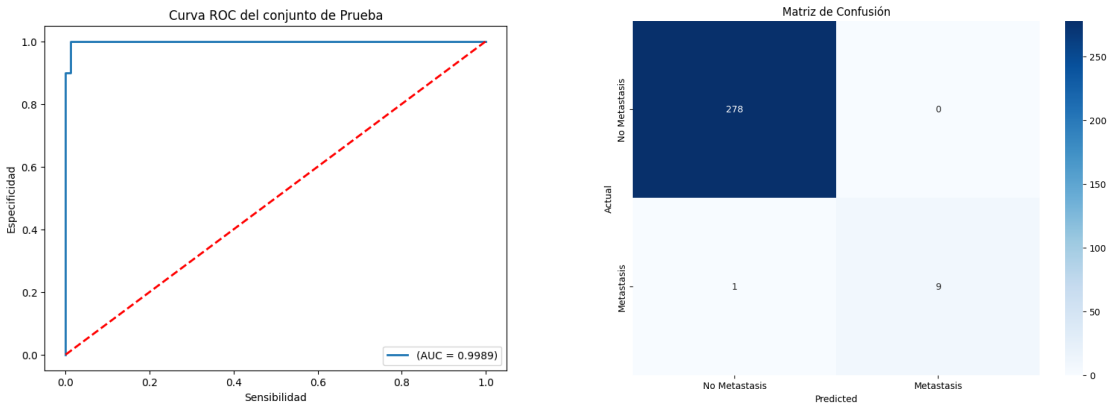


(a) Curva ROC (b) Matriz de confusión

Figura 5.15: Resultados del modelo LR-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.99	0.99
Metástasis	0.93	1.00	0.96

Tabla 5.15: Métricas del modelo LR-ADA para cáncer de próstata



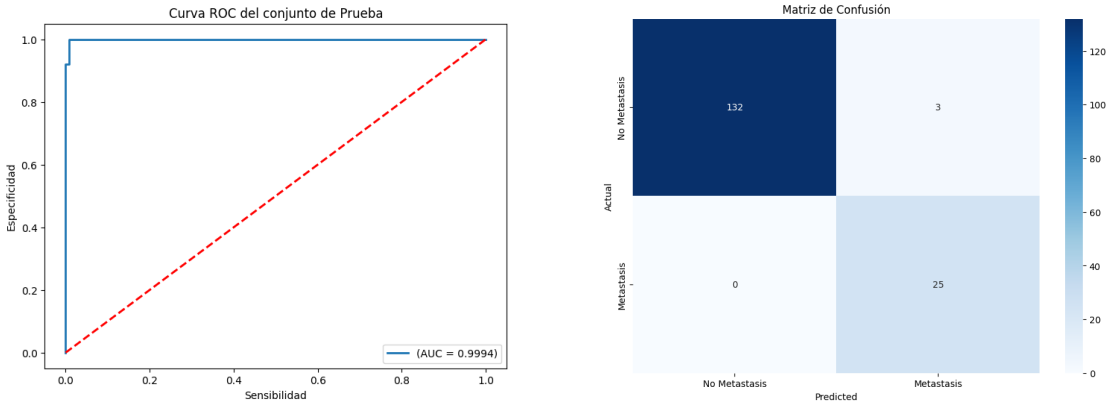
(a) Curva ROC (b) Matriz de confusión

Figura 5.16: Resultados del modelo LR-ADA para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.16: Métricas del modelo LR-ADA para cáncer de mama

5.3.1.3. Conjunto de datos GAN

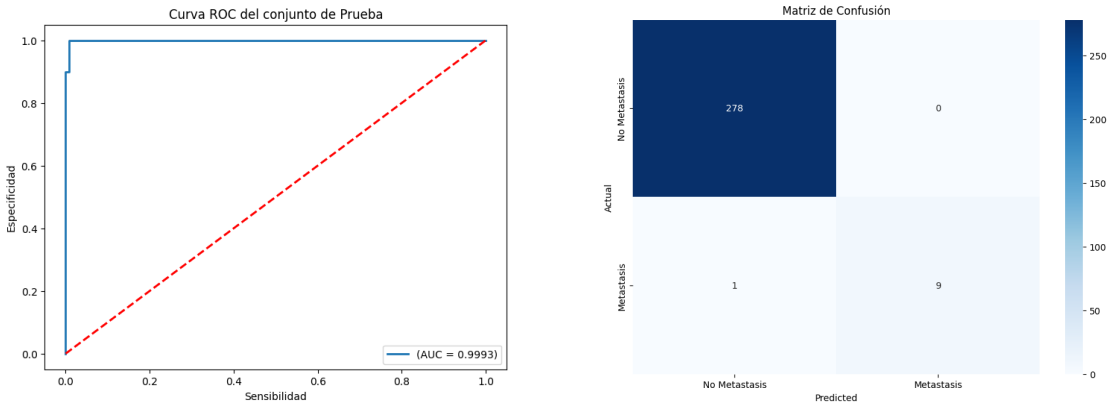


(a) Curva ROC (b) Matriz de confusión

Figura 5.17: Resultados del modelo LR-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.17: Métricas del modelo LR-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.18: Resultados del modelo LR-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.18: Métricas del modelo LR-GAN para cáncer de mama

5.4. Random Forest

En el contexto del modelo de Random Forest (RF), los hiperparámetros utilizados en el proceso de entrenamiento son los siguientes:

- **n_estimators**: Se ha utilizado un valor único de 100 para el número de árboles en el bosque.
- **max_depth**: Se han considerado valores de 10, 20 y 30 para la profundidad máxima de los árboles.
- **min_samples_split**: Se han considerado valores de 5 y 10 para el número mínimo de muestras requeridas para dividir un nodo interno.
- **min_samples_leaf**: Se han considerado valores de 2 y 4 para el número mínimo de muestras requeridas para estar en un nodo hoja.
- **bootstrap**: Se ha configurado como "True" para indicar si se utiliza el muestreo con reemplazo.

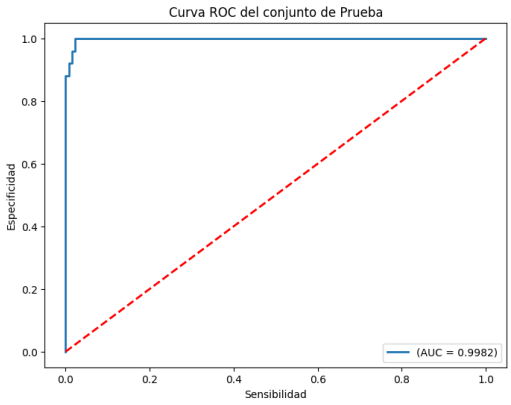
Por otro lado, en este caso se han juntado los conjuntos de entrenamiento y validación, y se ha realizado validación cruzada con KFold.

5.4.1. Evaluación del modelo

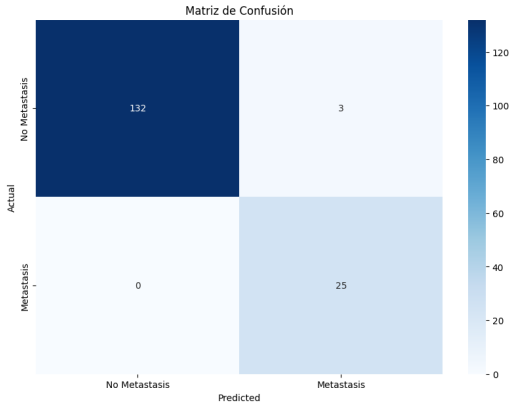
5.4.1.1. Conjunto de datos SMOTETomek

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.19: Métricas del modelo RF-ST para cáncer de próstata

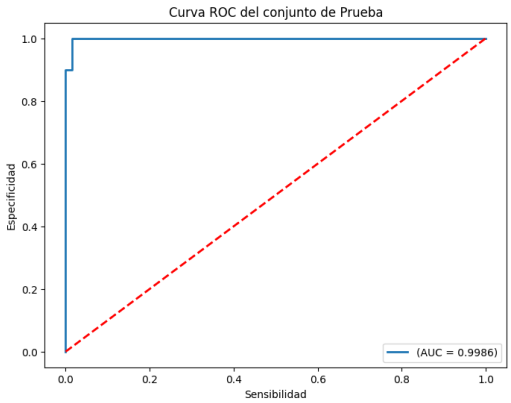


(a) Curva ROC

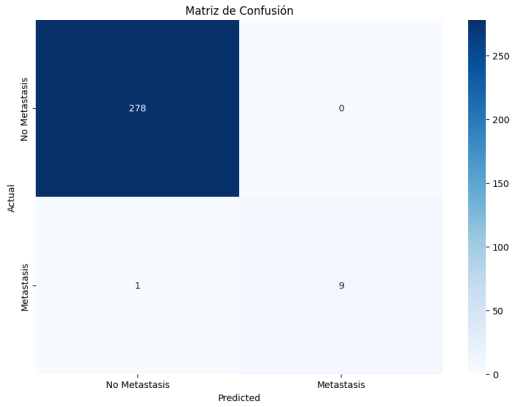


(b) Matriz de confusión

Figura 5.19: Resultados del modelo RF-ST para cáncer de próstata



(a) Curva ROC



(b) Matriz de confusión

Figura 5.20: Resultados del modelo RF-ST para cáncer de mama

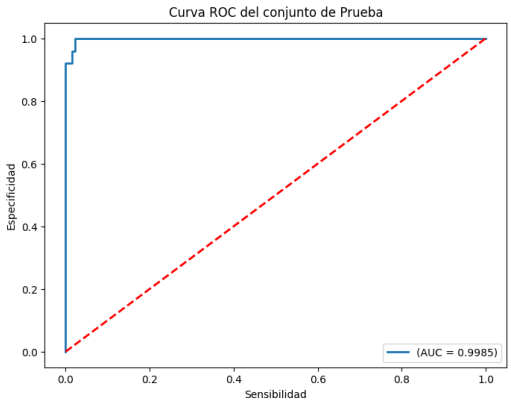
	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.20: Métricas del modelo RF-ST para cáncer de mama

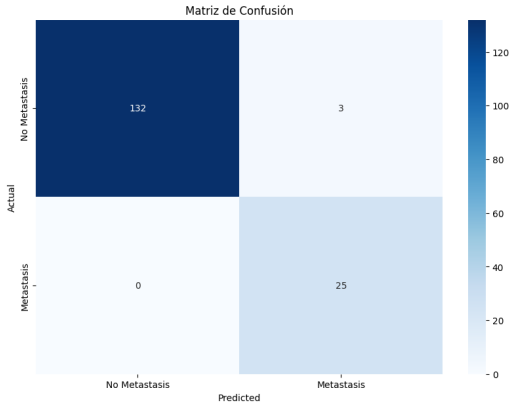
5.4.1.2. Conjunto de datos ADASYN

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.21: Métricas del modelo RF-ADA para cáncer de próstata

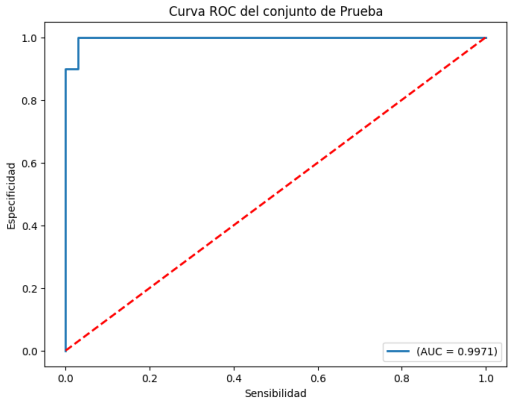


(a) Curva ROC

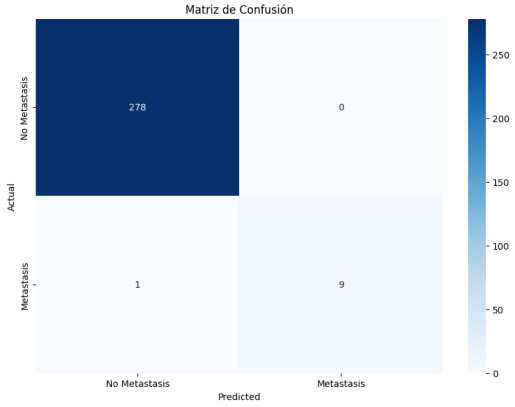


(b) Matriz de confusión

Figura 5.21: Resultados del modelo RF-ADA para cáncer de próstata



(a) Curva ROC



(b) Matriz de confusión

Figura 5.22: Resultados del modelo RF-ADA para cáncer de mama

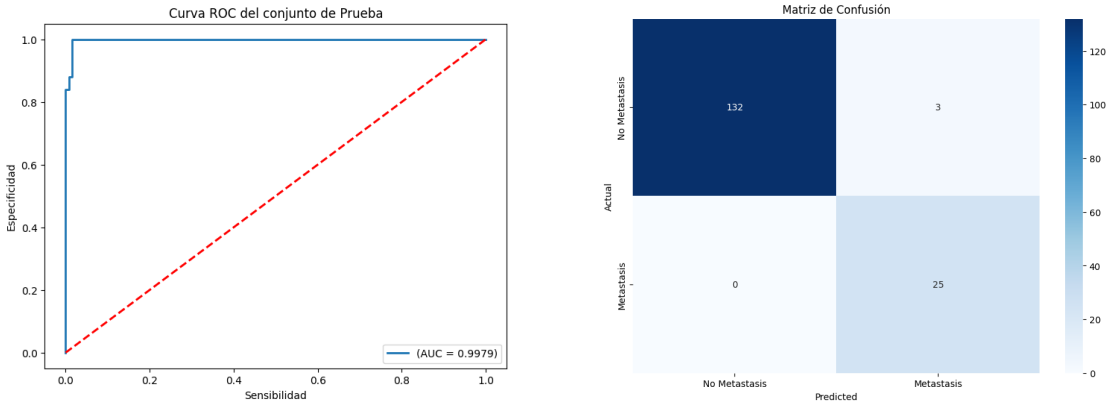
	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.22: Métricas del modelo RF-ADA para cáncer de mama

5.4.1.3. Conjunto de datos GAN

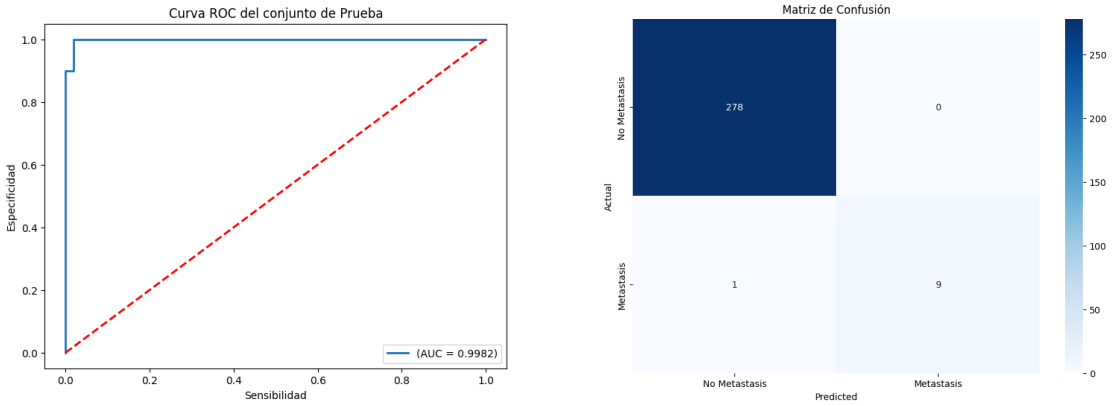
	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.23: Métricas del modelo RF-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.23: Resultados del modelo RF-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.24: Resultados del modelo RF-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.24: Métricas del modelo RF-GAN para cáncer de mama

5.5. Gradient boosting: XGBoost

Para el proceso de entrenamiento del modelo de Extreme Gradient Boosting (XGBoost) se han utilizado los siguientes hiperparámetros:

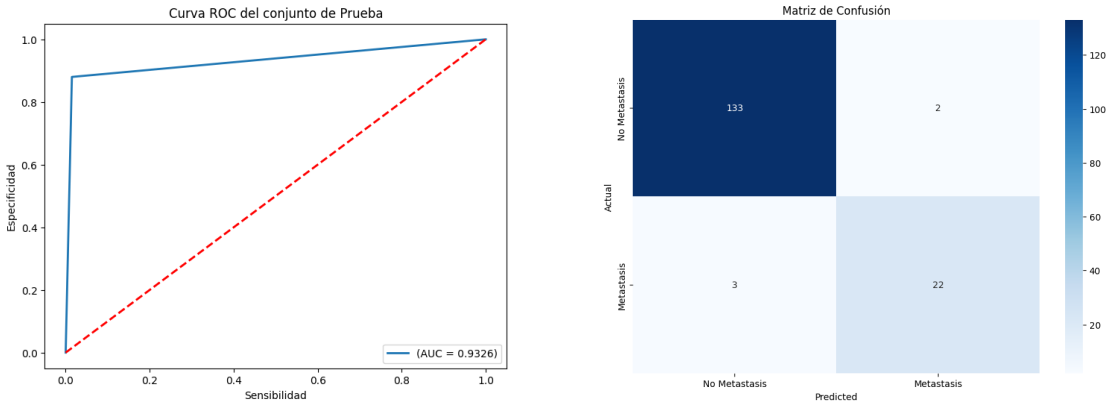
- **objective**: Se utiliza la función de pérdida "binary:logistic".
- **max_depth**: Se ha definido como un número entero en el rango de 3 a 5.
- **eta**: Se ha definido como número real en un rango logarítmico de 0.01 a 0.3.

- **eval_metric**: Se utiliza "logloss" como métrica de evaluación.

En este caso, se ha entrenado el modelo con el conjunto de entrenamiento, usando el conjunto de validación para validarlo, y se ha realizado optimización de hiperparámetros con Optuna.

5.5.1. Evaluación del modelo

5.5.1.1. Conjunto de datos SMOTETomek



(a) Curva ROC (b) Matriz de confusión

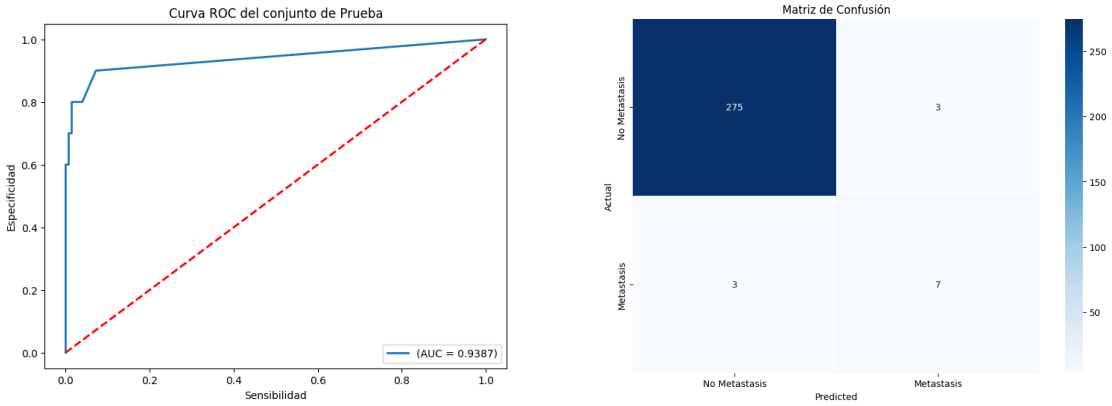
Figura 5.25: Resultados del modelo XGB-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.98	0.99	0.98
Metástasis	0.92	0.88	0.90

Tabla 5.25: Métricas del modelo XGB-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.99	0.99	0.99
Metástasis	0.70	0.70	0.70

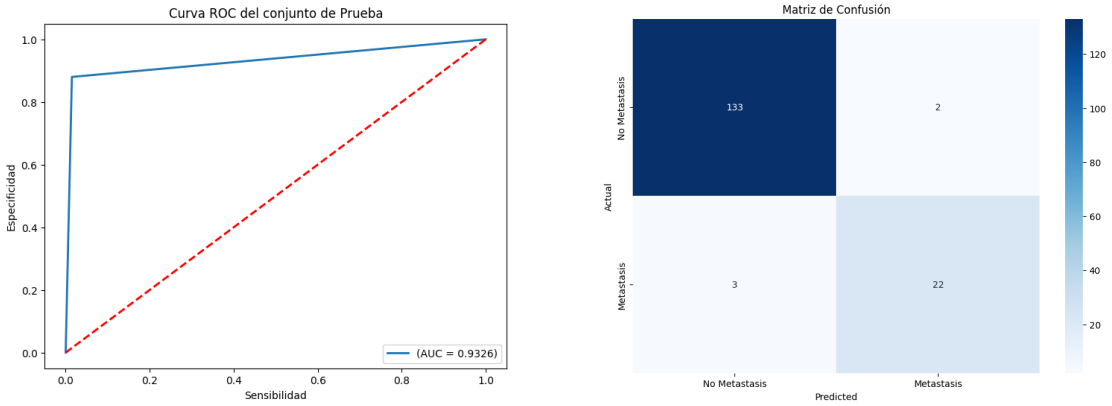
Tabla 5.26: Métricas del modelo XGB-ST para cáncer de mama



(a) Curva ROC (b) Matriz de confusión

Figura 5.26: Resultados del modelo XGB-ST para cáncer de mama

5.5.1.2. Conjunto de datos ADASYN



(a) Curva ROC (b) Matriz de confusión

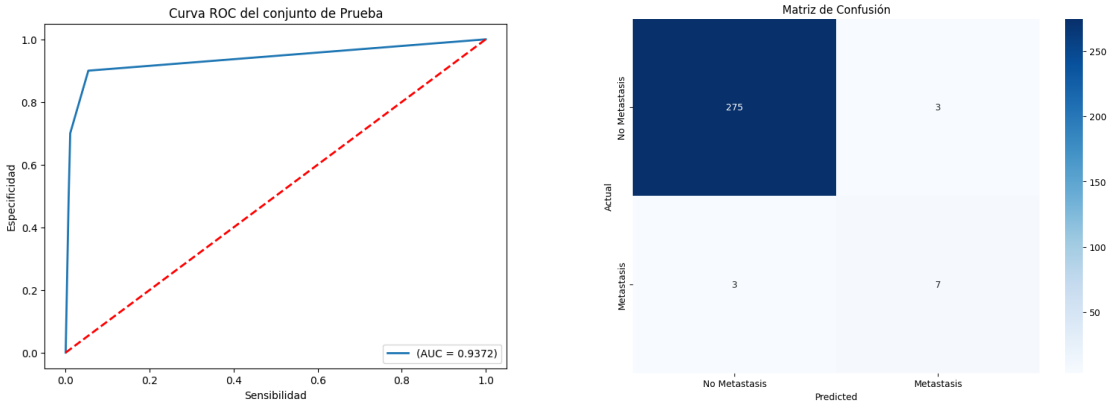
Figura 5.27: Resultados del modelo XGB-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.98	0.99	0.98
Metástasis	0.92	0.88	0.90

Tabla 5.27: Métricas del modelo XGB-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.99	0.99	0.99
Metástasis	0.70	0.70	0.70

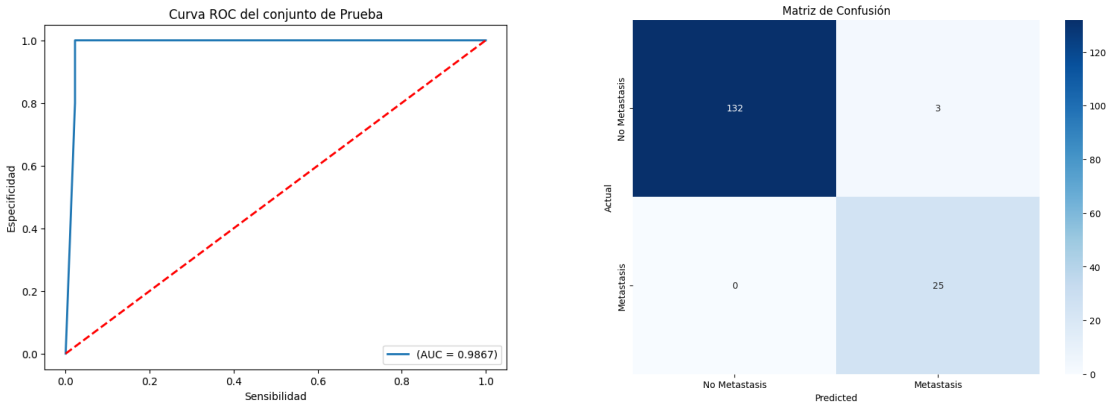
Tabla 5.28: Métricas del modelo XGB-ADA para cáncer de mama



(a) Curva ROC (b) Matriz de confusión

Figura 5.28: Resultados del modelo XGB-ADA para cáncer de mama

5.5.1.3. Conjunto de datos GAN



(a) Curva ROC (b) Matriz de confusión

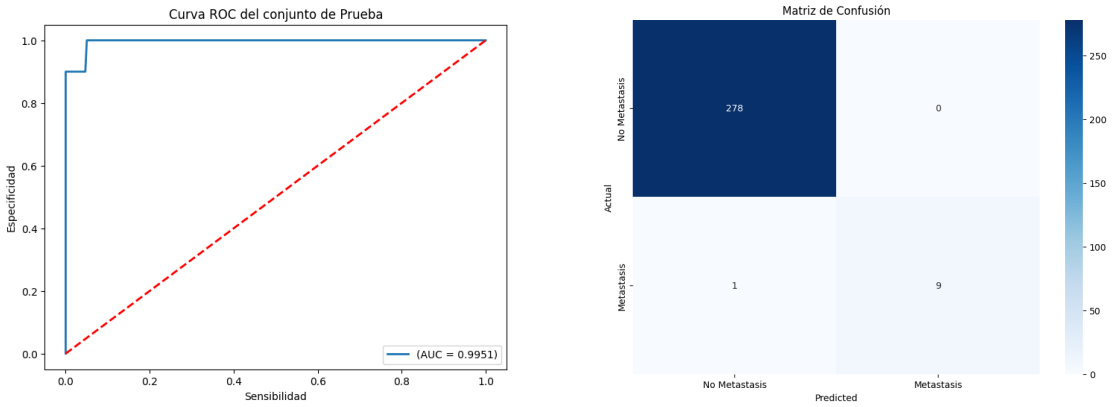
Figura 5.29: Resultados del modelo XGB-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.29: Métricas del modelo XGB-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.30: Métricas del modelo XGB-GAN para cáncer de mama



(a) Curva ROC (b) Matriz de confusión

Figura 5.30: Resultados del modelo XGB-GAN para cáncer de mama

5.6. Deep Neural Network

Para el proceso de entrenamiento de la Red Neuronal Profunda se han utilizado los siguientes hiperparámetros:

- **dropout_rate**: Es el porcentaje de neuronas que se desconectan de manera aleatoria en una capa durante cada iteración. Los valores utilizados son números reales en el rango de 0.1 y 0.5.
- **units**: Es la cantidad de neuronas por capa. Para cada capa se ha definido un número entero entre la anterior y la siguiente capa.

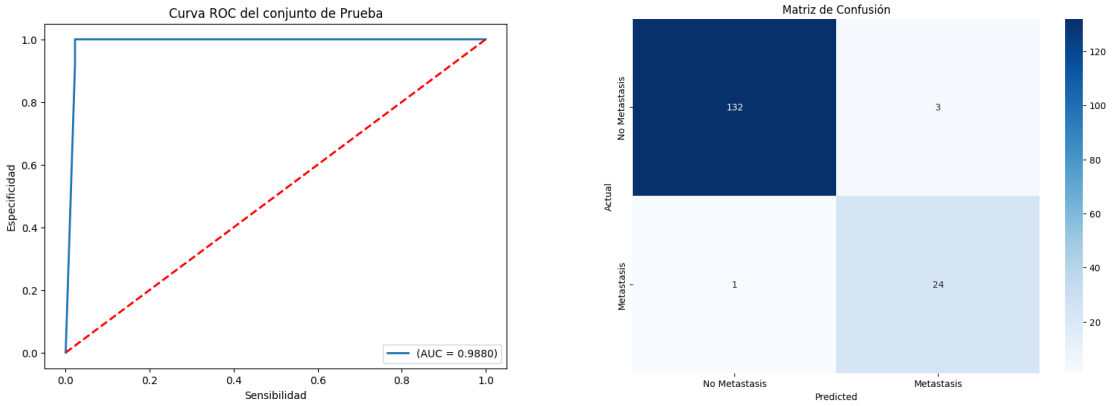
En este caso, se ha entrenado el modelo con el conjunto de entrenamiento, usando el conjunto de validación para validarlo, y se ha realizado optimización de hiperparámetros con Optuna.

5.6.1. Evaluación del modelo

5.6.1.1. Conjunto de datos SMOTETomek

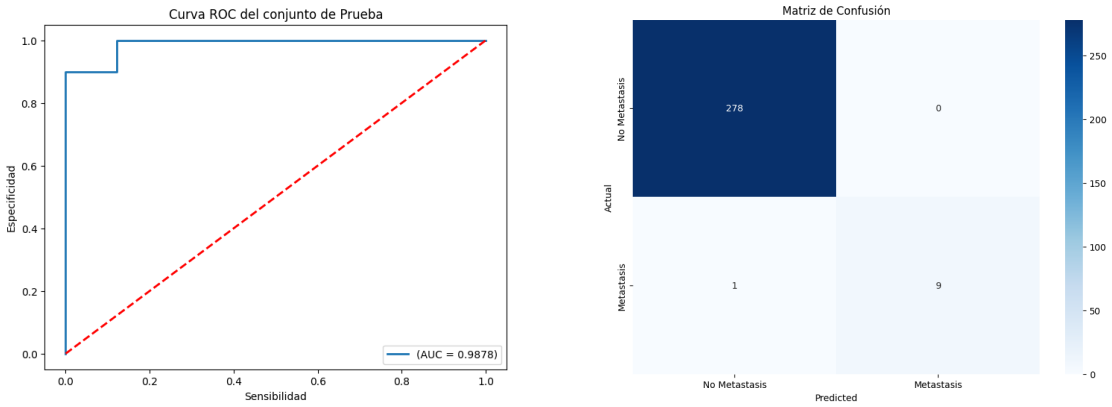
	Precision	Recall	F1-score
No Metástasis	0.99	0.98	0.99
Metástasis	0.89	0.96	0.92

Tabla 5.31: Métricas del modelo DNN-ST para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.31: Resultados del modelo DNN-ST para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.32: Resultados del modelo DNN-ST para cáncer de mama

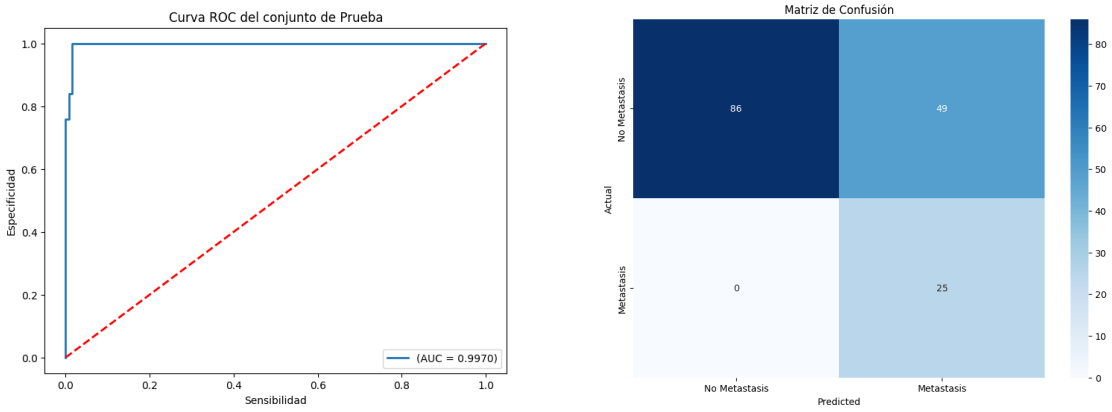
	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.32: Métricas del modelo DNN-ST para cáncer de mama

5.6.1.2. Conjunto de datos ADASYN

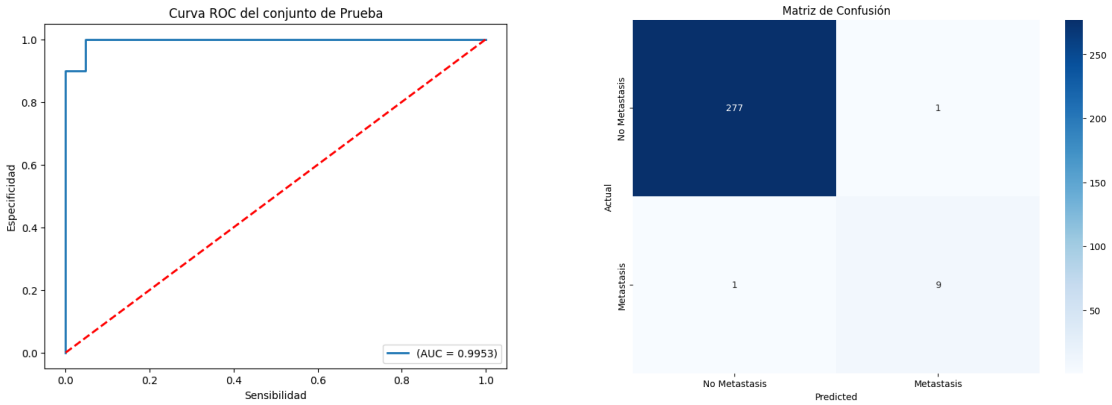
	Precision	Recall	F1-score
No Metástasis	1.00	0.64	0.78
Metástasis	0.34	1.00	0.51

Tabla 5.33: Métricas del modelo DNN-ADA para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.33: Resultados del modelo DNN-ADA para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.34: Resultados del modelo DNN-ADA para cáncer de mama

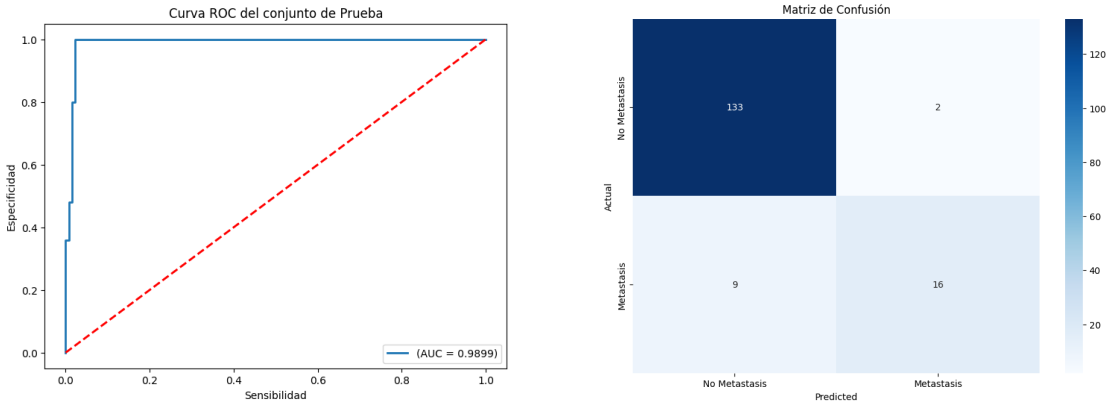
	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	0.90	0.90	0.90

Tabla 5.34: Métricas del modelo DNN-ADA para cáncer de mama

5.6.1.3. Conjunto de datos GAN

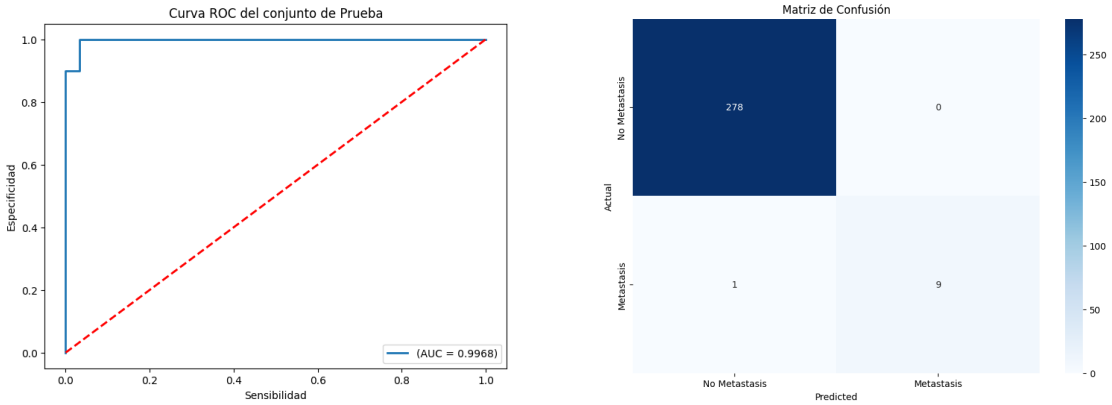
	Precision	Recall	F1-score
No Metástasis	0.94	0.99	0.96
Metástasis	0.89	0.64	0.74

Tabla 5.35: Métricas del modelo DNN-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.35: Resultados del modelo DNN-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.36: Resultados del modelo DNN-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

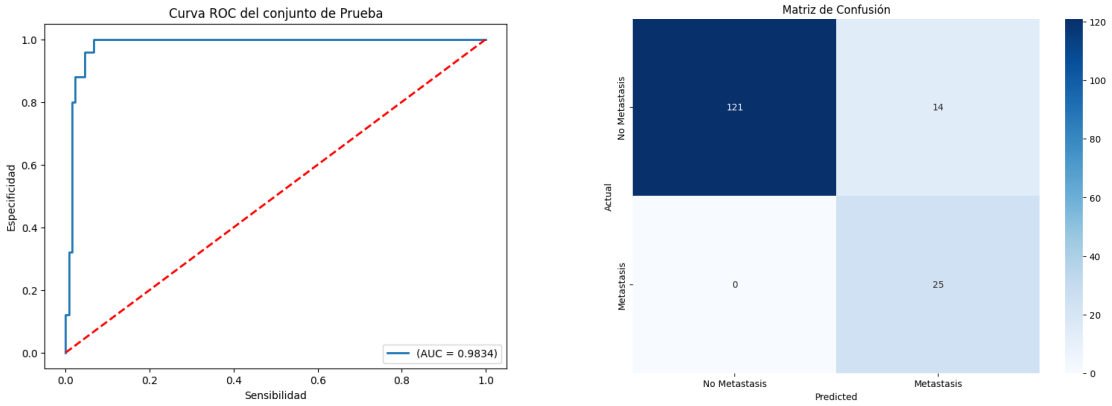
Tabla 5.36: Métricas del modelo DNN-GAN para cáncer de mama

5.7. TabNetClassifier

Para el entrenamiento del modelo TabNetClassifier se ha decidido utilizar los hiperparámetros predeterminados, ya que ofrecen mejores resultados. Al igual que en el caso anterior, se ha entrenado el modelo con el conjunto de entrenamiento, usando el conjunto de validación para validarlo.

5.7.1. Evaluación del modelo

5.7.1.1. Conjunto de datos SMOTETomek

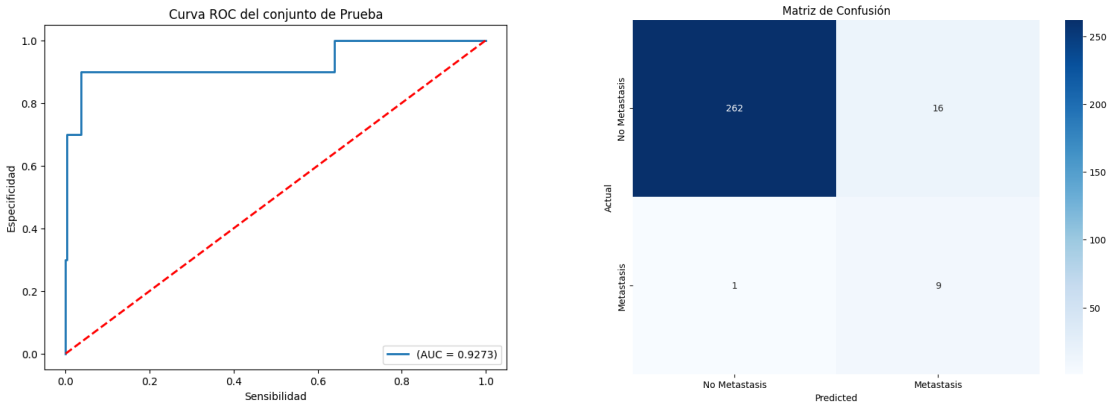


(a) Curva ROC (b) Matriz de confusión

Figura 5.37: Resultados del modelo TABNET-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.90	0.95
Metástasis	0.64	1.00	0.78

Tabla 5.37: Métricas del modelo TABNET-ST para cáncer de próstata



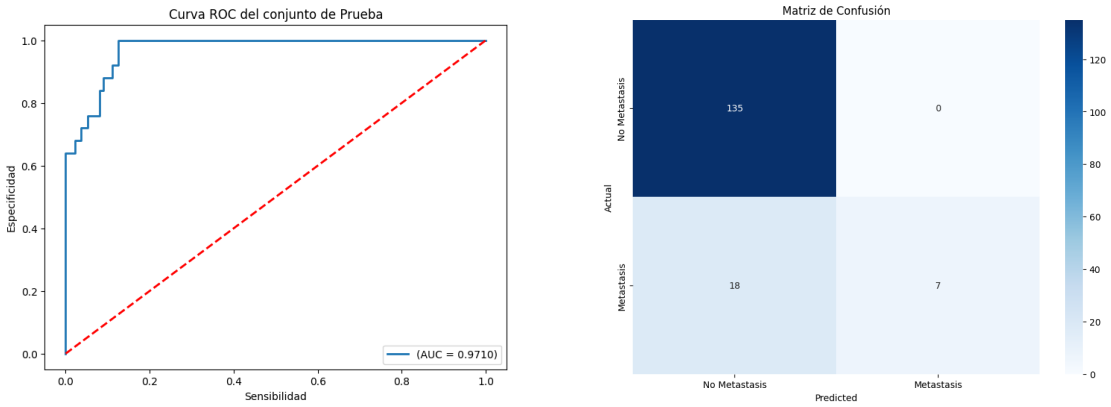
(a) Curva ROC (b) Matriz de confusión

Figura 5.38: Resultados del modelo TABNET-ST para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	0.94	0.97
Metástasis	0.36	0.90	0.51

Tabla 5.38: Métricas del modelo TABNET-ST para cáncer de mama

5.7.1.2. Conjunto de datos ADASYN

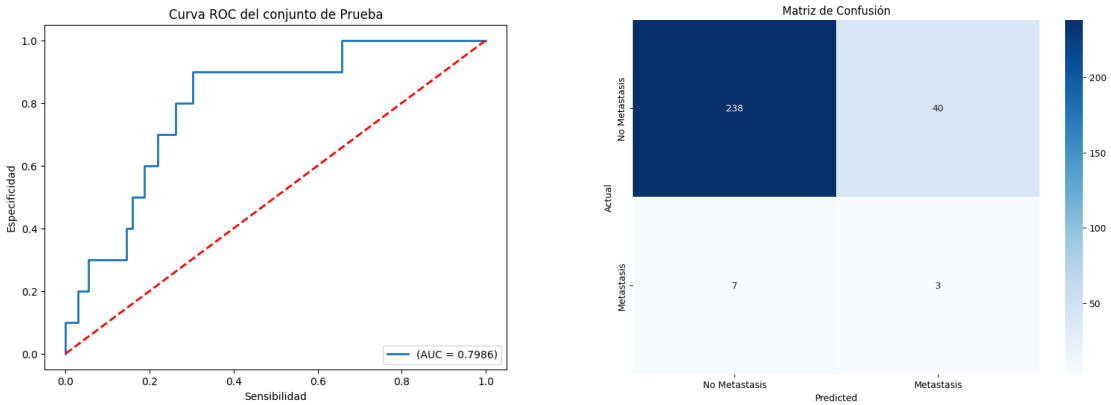


(a) Curva ROC (b) Matriz de confusión

Figura 5.39: Resultados del modelo TABNET-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.88	1.00	0.94
Metástasis	1.00	0.28	0.44

Tabla 5.39: Métricas del modelo TABNET-ADA para cáncer de próstata



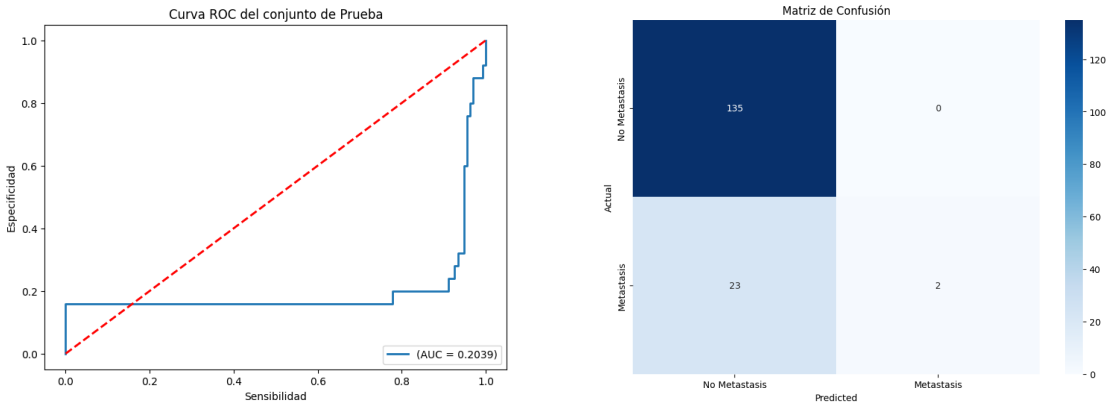
(a) Curva ROC (b) Matriz de confusión

Figura 5.40: Resultados del modelo TABNET-ADA para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	0.97	0.86	0.91
Metástasis	0.07	0.30	0.11

Tabla 5.40: Métricas del modelo TABNET-ADA para cáncer de mama

5.7.1.3. Conjunto de datos GAN

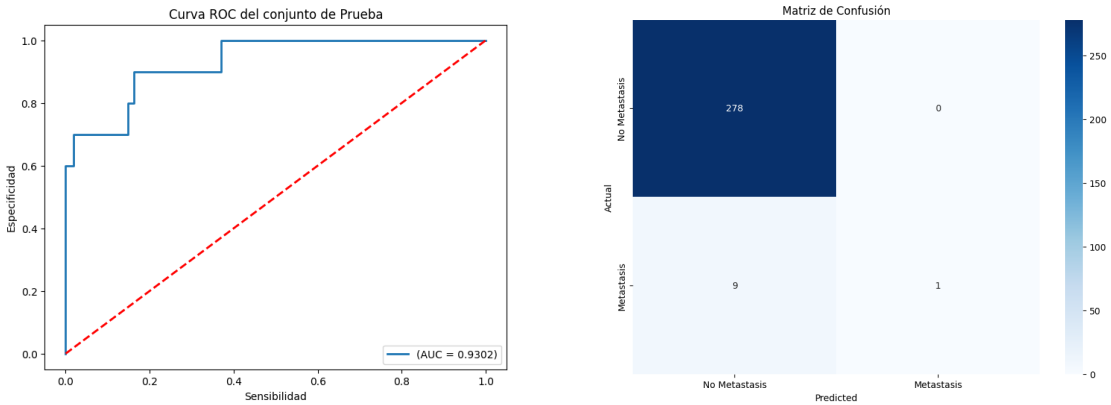


(a) Curva ROC (b) Matriz de confusión

Figura 5.41: Resultados del modelo TABNET-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	0.85	1.00	0.92
Metástasis	1.00	0.08	0.15

Tabla 5.41: Métricas del modelo TABNET-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.42: Resultados del modelo TABNET-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	0.97	1.00	0.98
Metástasis	1.00	0.10	0.18

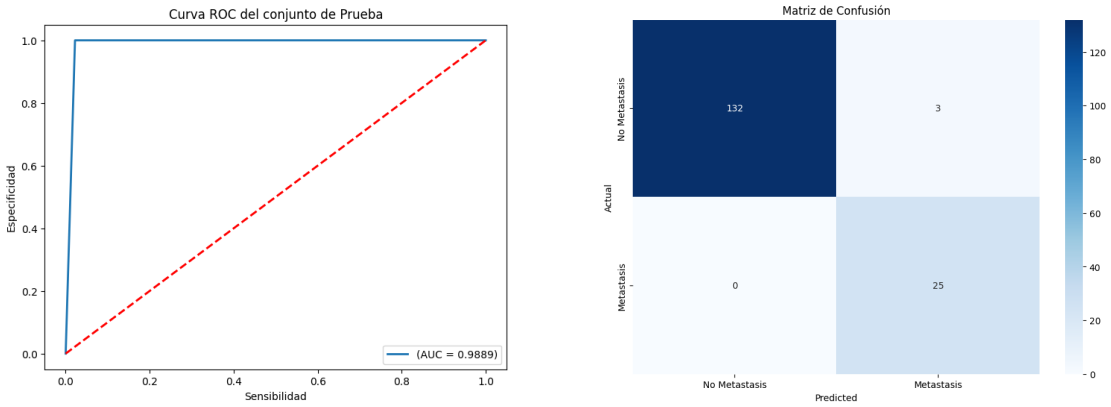
Tabla 5.42: Métricas del modelo TABNET-GAN para cáncer de mama

5.8. AutoGluon: TabularPredictor

En este caso se utiliza un ensamble de modelos para mejorar la precisión. Además, este paquete automatiza la optimización de hiperparámetros y la selección de modelos. El entrenamiento se ha realizado con el conjunto de entrenamiento, usando el conjunto de validación para validarlo.

5.8.1. Evaluación del modelo

5.8.1.1. Conjunto de datos SMOTETomek

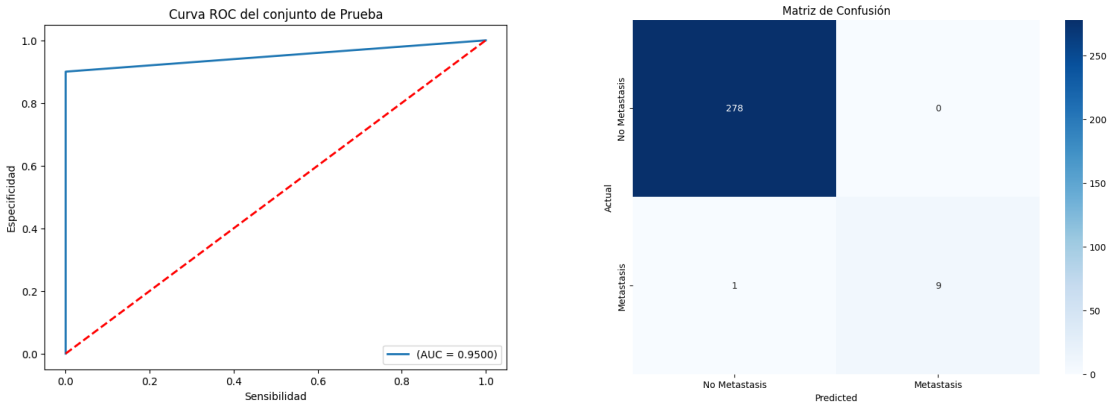


(a) Curva ROC (b) Matriz de confusión

Figura 5.43: Resultados del modelo TP-ST para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.43: Métricas del modelo TP-ST para cáncer de próstata



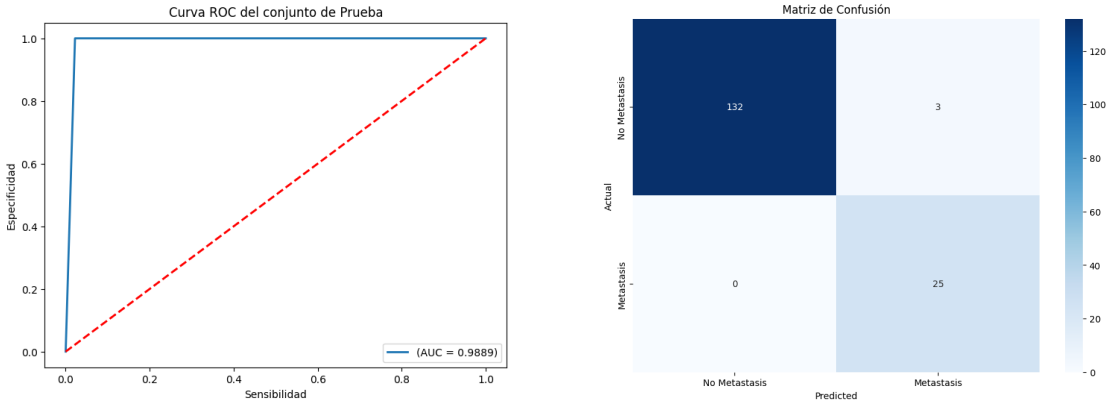
(a) Curva ROC (b) Matriz de confusión

Figura 5.44: Resultados del modelo TP-ST para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.44: Métricas del modelo TP-ST para cáncer de mama

5.8.1.2. Conjunto de datos ADASYN

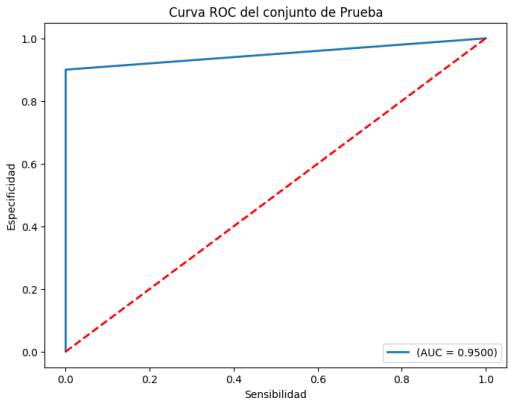


(a) Curva ROC (b) Matriz de confusión

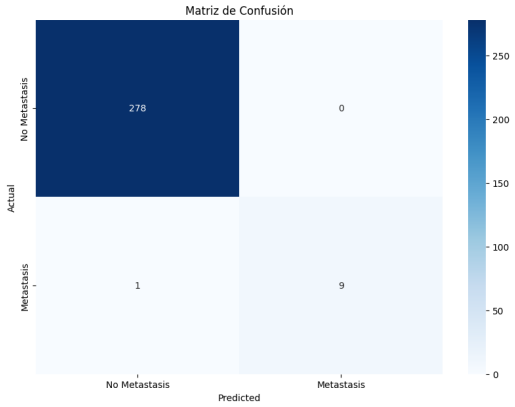
Figura 5.45: Resultados del modelo TP-ADA para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.45: Métricas del modelo TP-ADA para cáncer de próstata



(a) Curva ROC



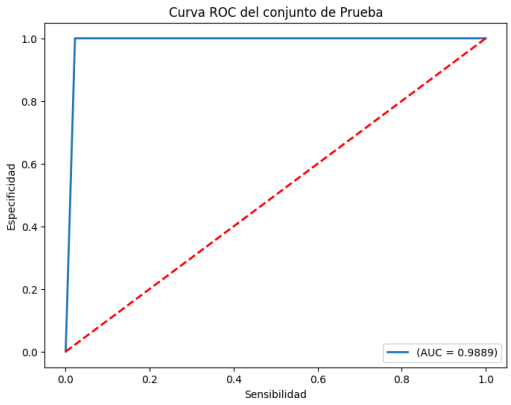
(b) Matriz de confusión

Figura 5.46: Resultados del modelo TP-ADA para cáncer de mama

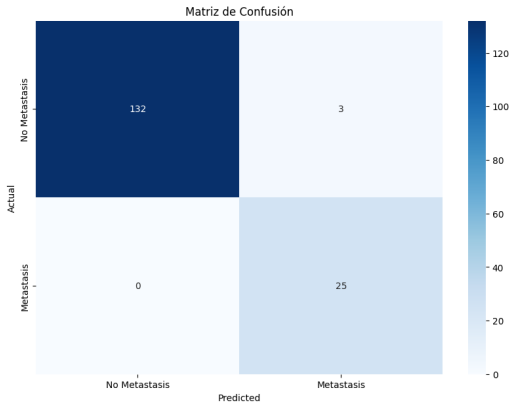
	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.46: Métricas del modelo TP-ADA para cáncer de mama

5.8.1.3. Conjunto de datos GAN



(a) Curva ROC

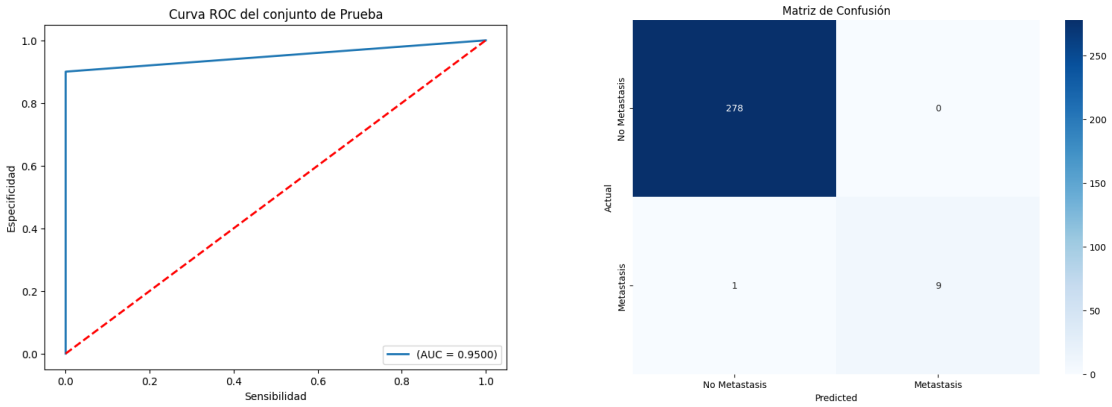


(b) Matriz de confusión

Figura 5.47: Resultados del modelo TP-GAN para cáncer de próstata

	Precision	Recall	F1-score
No Metástasis	1.00	0.98	0.99
Metástasis	0.89	1.00	0.94

Tabla 5.47: Métricas del modelo TP-GAN para cáncer de próstata



(a) Curva ROC (b) Matriz de confusión

Figura 5.48: Resultados del modelo TP-GAN para cáncer de mama

	Precision	Recall	F1-score
No Metástasis	1.00	1.00	1.00
Metástasis	1.00	0.90	0.95

Tabla 5.48: Métricas del modelo TP-GAN para cáncer de mama

5.9. Análisis de influencia de genes

Al comparar los resultados de cada modelo, se ha observado que los modelos SVC (Support Vector Classifier), LR (Logistic Regression), RF (Random Forest) y AutoGluon son los que proporcionan los resultados más óptimos. Entre ellos, AutoGluon destaca por ofrecer el mayor nivel de confianza en las predicciones, debido a su capacidad para ensamblar múltiples modelos, lo que le permite evitar el sobreajuste y mejorar la precisión de las predicciones.

Sin embargo, a pesar de sus ventajas, el uso de AutoGluon para el análisis de influencia de genes en la detección de casos de metástasis presenta un desafío significativo debido a su elevado coste computacional. Además, la extracción de coeficientes con AutoGluon se estima que durará alrededor de 5 días. AutoGluon no proporciona las mismas herramientas de interpretación y análisis de modelos que ofrece scikit-learn, lo que complica su implementación para estudios detallados de influencia de genes. De manera similar, el modelo SVC también resulta ser computacionalmente costoso para el análisis detallado de influencia de genes.

Por estos motivos, se ha decidido utilizar el modelo de Regresión Logística (LR) para el análisis de influencia de genes. La Regresión Logística, aunque pueda no ser tan compleja como AutoGluon o SVC, ofrece un equilibrio adecuado entre rendimiento y coste computacional. Además, proporciona herramientas claras y eficaces para la interpretación de los resultados, lo que la convierte en una opción viable y práctica para el análisis de influencia de genes en la detección de metástasis.

En las imágenes 5.49 y 5.50 se observan los genes que mayor influencia tienen en la predicción de la metástasis en cáncer de próstata y cáncer de mama respectivamente. Los genes con coeficientes positivos están asociados con un aumento en la probabilidad de metástasis cuando su expresión aumenta, mientras que los genes con coeficientes negativos están asociados con una disminución en la probabilidad de metástasis cuando su expresión aumenta.

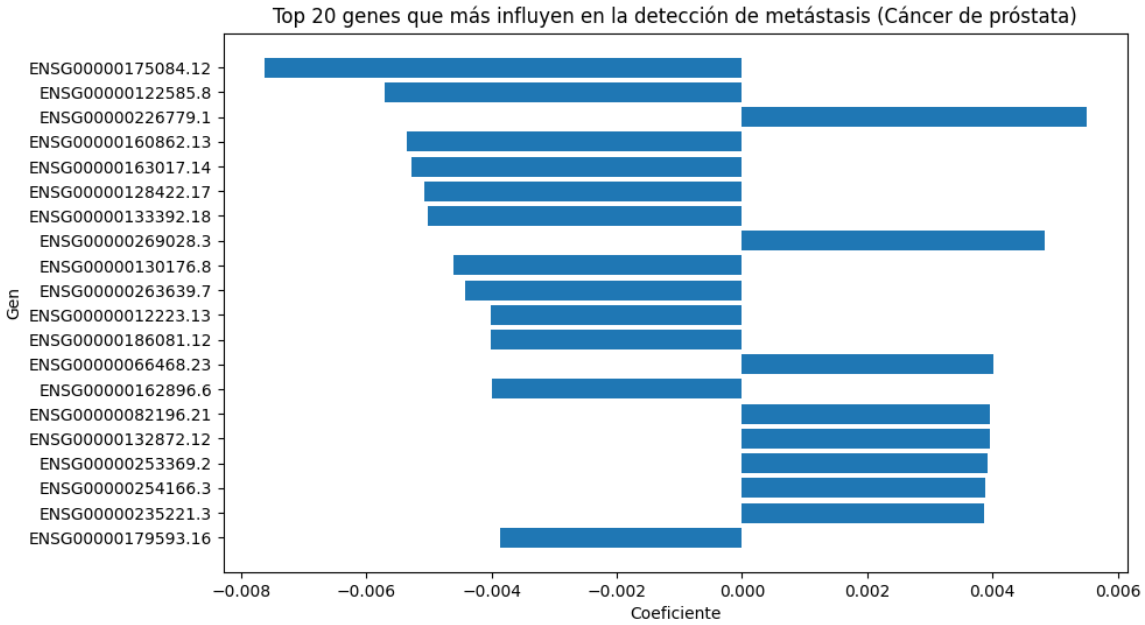


Figura 5.49: Genes que más influyen en el cáncer de próstata

Al filtrar el análisis para incluir solo los genes con coeficientes positivos, se pueden identificar aquellos que aumentan la probabilidad de metástasis cuando su expresión aumenta. Este enfoque permite centrarse en los factores genéticos que potencian la progresión metastásica. Los resultados de este análisis se presentan en las imágenes 5.51 y 5.52.

Es importante destacar que los genes más influyentes varían significativamente entre los diferentes tipos de cáncer analizados. No se observa coincidencia entre los genes principales asociados con la metástasis en cada tipo de cáncer, lo que sugiere una especificidad única en los mecanismos genéticos que impulsan la metástasis en distintos contextos tumorales. Esta variabilidad subraya la necesidad de enfoques personalizados en la investigación y el tratamiento del cáncer metastásico.

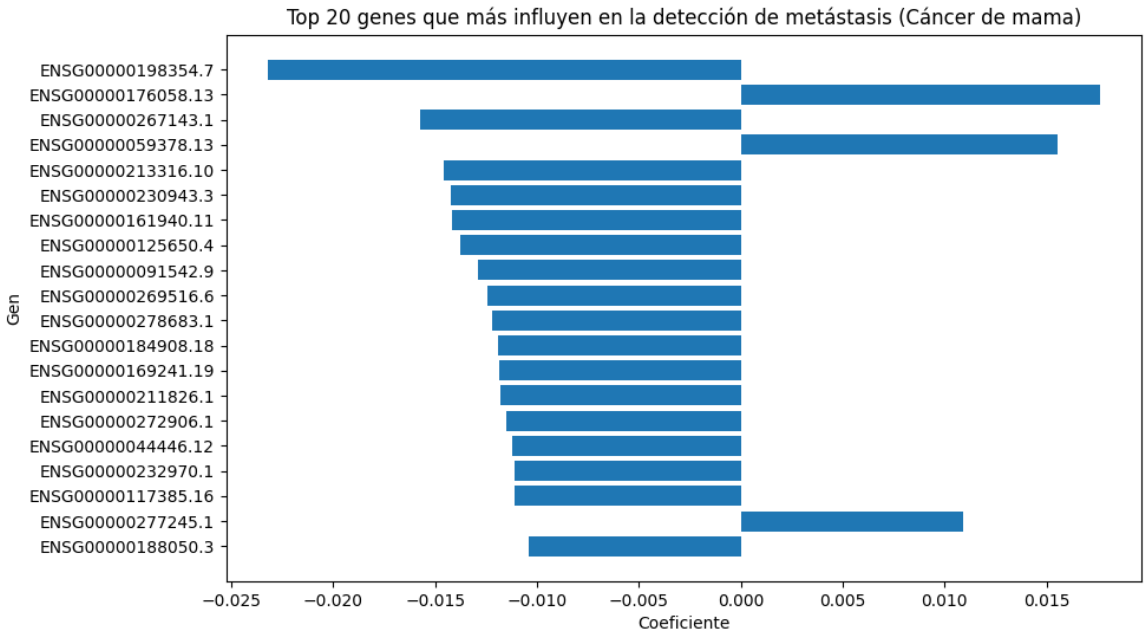


Figura 5.50: Genes que más influyen en el cáncer de mama

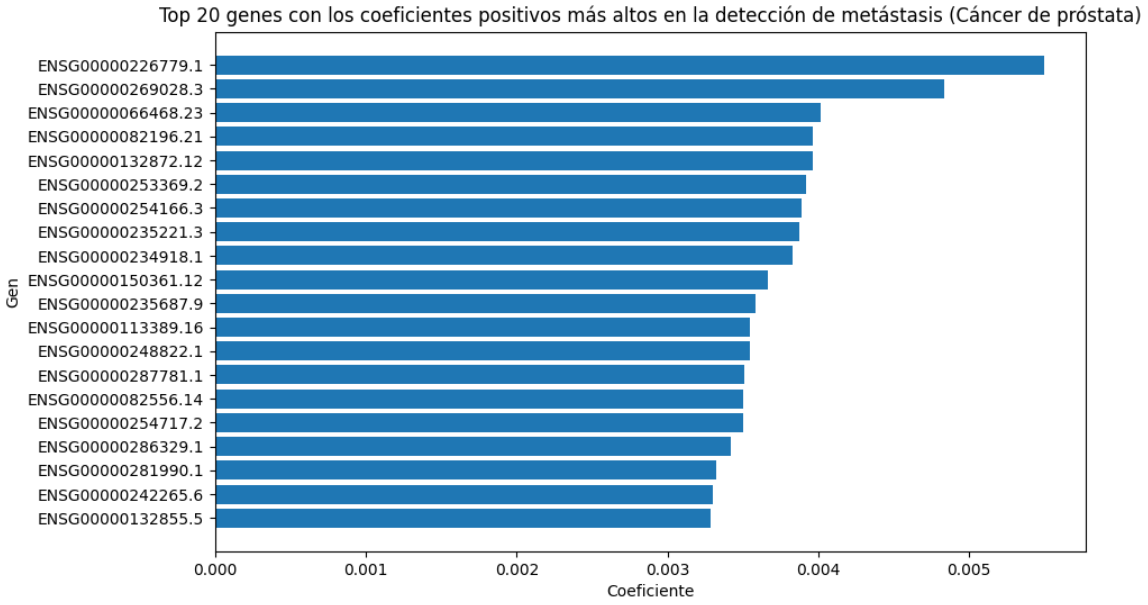


Figura 5.51: Genes con coeficiente positivo que más influyen en el cáncer de próstata

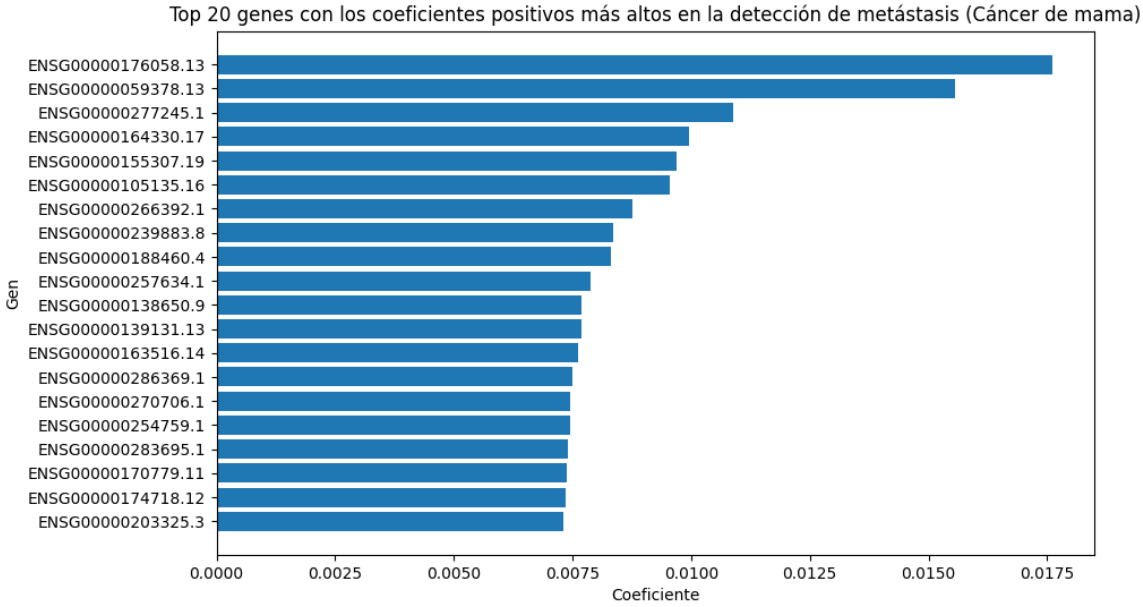


Figura 5.52: Genes con coeficiente positivo que más influyen en el cáncer de mama

Capítulo 6

Conclusiones y trabajos futuros

Este proyecto ha tenido como objetivo desarrollar modelos de machine learning para la detección de metástasis a partir de datos de expresión génica. Los modelos resultantes han demostrado ser eficientes para resolver el problema propuesto, y los resultados obtenidos respaldan el uso del machine learning en la detección de metástasis. Se han realizado evaluaciones exhaustivas utilizando diversas métricas para asegurar la robustez y precisión de los modelos desarrollados.

En conclusión, los modelos más óptimos identificados en este estudio para la detección de metástasis son el Support Vector Classifier (SVC), Random Forest (RF), Logistic Regression (LR) y AutoGluon. Estos modelos han demostrado un desempeño destacado en términos de f1-score y recall. En el caso del cáncer de mama, se ha logrado un recall de 0.9, mientras que para el cáncer de próstata se ha obtenido un recall perfecto de 1. Para ambos tipos de cáncer, el valor mínimo de f1-score ha sido de 0.94 para los modelos óptimos.

Además, el uso del modelo de Logistic Regression ha permitido identificar genes que podrían servir como biomarcadores, ofreciendo así nuevas oportunidades para el diagnóstico precoz y la personalización de los tratamientos. Estos hallazgos subrayan la eficacia de los enfoques de machine learning en la investigación del cáncer y su potencial para mejorar los resultados clínicos en pacientes con metástasis.

Para trabajos futuros, se pueden explorar varias líneas de investigación que amplíen y profundicen los hallazgos de este estudio. Además, el notebook y los datos necesarios para la utilización del código desarrollado han quedado disponibles de forma pública en el repositorio de GitHub, bajo una licencia Creative Commons. En primer lugar, se pueden desarrollar modelos adicionales para otros tipos de cáncer, lo que permitiría evaluar la eficacia de los métodos propuestos en un espectro más amplio.

Asimismo, se pueden crear modelos que mezclen diferentes tipos de cáncer, con el objetivo de identificar patrones genéticos comunes y diferenciales en la metástasis entre diversos tipos

de cáncer. Esta estrategia podría proporcionar una visión más integral de los mecanismos metastásicos.

Además, con la disponibilidad de una mayor capacidad computacional, se podría realizar un análisis exhaustivo de los genes significativos utilizando el modelo de AutoGluon. Este análisis permitiría identificar de manera más precisa los genes clave asociados con la metástasis.

Finalmente, investigar la función biológica de los genes más influyentes identificados podría ser crucial para comprender mejor su papel en la metástasis. Esta investigación podría revelar nuevos biomarcadores, ofreciendo oportunidades para desarrollar tratamientos más efectivos y personalizados para combatir la propagación del cáncer.

Anexo A

Código para la creación del dataset

A.1. Extracción de datos

Este anexo incluye el código R utilizado para la descarga, preparación y almacenamiento de los datos de expresión génica desde TCGA.

```r
library(TCGAbiolinks)
# Funcion para crear una consulta y guardar los datos
geneDataTCGA <- function(proj, path, sample_type) {
  query <- GDCquery(project = proj,
                    data.category = 'Transcriptome Profiling',
                    data.type = 'Gene Expression Quantification',
                    experimental.strategy = 'RNA-Seq',
                    workflow.type = 'STAR - Counts',
                    access = 'open', sample.type = sample_type)
  GDCdownload(query, directory = path)
  rna <- GDCprepare(query, directory = path)
  write_rds(rna, file=paste0(path, "/tpm_", proj, ".RDS"))
}
# Funcion para extraer los datos
extractAllGenes <- function(data, metric) {
  tpmDat <- as.data.frame(assays(data)[[metric]])
  tpmGene <- as.data.frame(t(tpmDat)) %>%
    rownames_to_column("barcode")
  return(tpmGene)
}
```

Programa A.1: Funciones para la extracción de datos

```r
# Descargar datos
geneDataTCGA('TCGA-PRAD', download_dir_TP, c("Primary Tumor"))
geneDataTCGA('TCGA-PRAD', download_dir_TM, c("Metastatic"))
geneDataTCGA('CMI-MPC', download_dir_TP, c("Primary Tumor"))
geneDataTCGA('WCDT-MCRPC', download_dir_TM, c("Metastatic"))
TM_genes_dat <- data.frame()
TP_genes_dat <- data.frame()
# Proceso para tumores primarios
tpm <- read_rds(paste0(download_dir_TP, "/tpm_TCGA-PRAD", ".RDS"))
TP_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TP_genes_temp, file="PRAD_data/PRAD-TP_genes_TCGA-PRAD.csv")
TP_genes_dat <- bind_rows(TP_genes_dat, TP_genes_temp)
tpm <- read_rds(paste0(download_dir_TP, "/tpm_CMI-MPC", ".RDS"))
TP_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TP_genes_temp, file="PRAD_data/PRAD-TP_genes-CMI-MPC.csv")
TP_genes_dat <- bind_rows(TP_genes_dat, TP_genes_temp)
write_csv(TP_genes_dat, "PRAD_data/PRAD-TP_genes.csv")
# Proceso para tumores metastaticos
tpm <- read_rds(paste0(download_dir_TM, "/tpm_TCGA-PRAD", ".RDS"))
TM_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TM_genes_temp, file="PRAD_data/PRAD-TM_genes_TCGA-PRAD.csv")
TM_genes_dat <- bind_rows(TM_genes_dat, TM_genes_temp)
tpm <- read_rds(paste0(download_dir_TM, "/tpm_WCDT-MCRPC", ".RDS"))
TM_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TM_genes_temp, file="PRAD_data/PRAD-TM_genes_WCDT-MCRPC.csv"
    )
TM_genes_dat <- bind_rows(TM_genes_dat, TM_genes_temp)
write_csv(TM_genes_dat, "PRAD_data/PRAD-TM_genes.csv")
```

Programa A.2: Creación de los archivos para el cáncer de próstata

```r
# Descargar datos
geneDataTCGA('TCGA-BRCA', download_dir_TP, c("Primary Tumor"))
geneDataTCGA('TCGA-BRCA', download_dir_TM, c("Metastatic"))
geneDataTCGA('CMI-MBC', download_dir_TM, c("Metastatic"))
geneDataTCGA('CMI-ASC', download_dir_TM, c("Metastatic"))
TM_genes_dat <- data.frame()
TP_genes_dat <- data.frame()
# Proceso para tumores primarios
```

```
tpm <- read_rds(paste0(download_dir_TP, "/tpm_TCGA-BRCA", ".RDS"))
TP_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TP_genes_temp, file="BRCA_data/BRCA-TP_genes.csv")
# Proceso para tumores metastaticos
tpm <- read_rds(paste0(download_dir_TM, "/tpm_TCGA-BRCA", ".RDS"))
TM_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TM_genes_temp, file="BRCA_data/BRCA-TM_genes_TCGA-BRCA.csv")
TM_genes_dat <- bind_rows(TM_genes_dat, TM_genes_temp)
tpm <- read_rds(paste0(download_dir_TM, "/tpm_CMI-MBC", ".RDS"))
TM_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TM_genes_temp, file="BRCA_data/BRCA-TM_genes_CMI-MBC.csv")
TM_genes_dat <- bind_rows(TM_genes_dat, TM_genes_temp)
tpm <- read_rds(paste0(download_dir_TM, "/tpm_CMI-ASC", ".RDS"))
TM_genes_temp <- extractAllGenes(tpm, "tpm_unstrand")
write_csv(TM_genes_temp, file="BRCA_data/BRCA-TM_genes_CMI-ASC.csv")
TM_genes_dat <- bind_rows(TM_genes_dat, TM_genes_temp)
write_csv(TM_genes_dat, "BRCA_data/BRCA-TM_genes.csv")
```

Programa A.3: Creación de los archivos para el cáncer de mama

A.2. Limpieza y normalización de datos

En este anexo se encuentra el código utilizado para realizar la limpieza y normalización de los datos recopilados.

```
def filter_and_normalize(data, expression_threshold=1, sample_
    threshold=0.1):
     metastatic_column = data['metastatic']
     data_without_metastatic = data.drop(columns=['metastatic', '
    barcode'])
     # Separar datos de metastasis y no metastasis
     metastatic_data = data_without_metastatic[metastatic_column == 1]
     non_metastatic_data = data_without_metastatic[metastatic_column ==
     0]
     # Filtrar genes con baja expresion en datos de metastasis
     num_samples_metastatic = metastatic_data.shape[0]
     metastatic_filtered = metastatic_data.loc[:, (metastatic_data >
    expression_threshold).sum(axis=0) > sample_threshold * num_samples_
    metastatic]
```

```
    # Filtrar genes con baja expresion en datos de no metastasis
    num_samples_non_metastatic = non_metastatic_data.shape[0]
    non_metastatic_filtered = non_metastatic_data.loc[:, (non_
metastatic_data > expression_threshold).sum(axis=0) > sample_
threshold * num_samples_non_metastatic]
    # Obtener los nombres de los genes que cumplen con los criterios
en cualquiera de los dos grupos
    filtered_genes = list(set(metastatic_filtered.columns) | set(non_
metastatic_filtered.columns))
    # Filtrar el dataset original para mantener solo los genes
importantes
    data_filtered = data_without_metastatic[filtered_genes]
    # Normalizacion logaritmica
    data_normalized = data_filtered.applymap(lambda x: np.log2(x + 1))
    data_normalized.insert(0, 'metastatic', metastatic_column)
    return data_normalized
```

Programa A.4: Función para filtrar y normalizar los datos

A.3. Creación de datos sintéticos mediante GAN

En este anexo se encuentra el código utilizado para la creación de datos sintéticos mediante GAN.

```
generator = build_generator(input_dim, data_dim)
discriminator = build_discriminator(data_dim)
gan = build_GAN(generator, discriminator)
for i in range(1,iterations+1):
    noise = np.random.normal(0,1,[batch_size,input_dim])
    batch_fake = generator.predict(noise)
    idx = np.random.randint(low=0, high=X_train_minority_np.shape[0],
size=batch_size)
    batch_real = X_train_minority_np[idx]
    discriminator.trainable = True
    d_loss_real = discriminator.train_on_batch(batch_real,real)
    d_loss_fake = discriminator.train_on_batch(batch_fake,fake)
    d_loss = 0.5 * np.add(d_loss_real, d_loss_fake)
    discriminator.trainable = False
    noise = np.random.normal(0, 1, (batch_size, input_dim))
```

```
15      g_loss = gan.train_on_batch(noise, np.ones(batch_size))
```

Programa A.5: Entrenamiento del modelo para generar datos sintéticos

Anexo B

Funciones auxiliares para la evaluación de los modelos

Este anexo inlcuye las funciones utilizadas para visualizar la evaluación de los modelos.

```
def plot_loss_acc(history):
    plt.plot(history.history['accuracy'], label='Train')
    plt.plot(history.history['val_accuracy'], label='Validation')
    plt.title('Model accuracy')
    plt.ylabel('Accuracy')
    plt.xlabel('Epoch')
    plt.legend()
    plt.show()
    plt.plot(history.history['loss'], label='Train')
    plt.plot(history.history['val_loss'], label='Validation')
    plt.title('Model loss')
    plt.ylabel('Loss')
    plt.xlabel('Epoch')
    plt.legend()
    plt.show()
```

Programa B.1: Función para visualizar las curvas de entrenamiento y evaluación de las redes neuronales

```
def plot_confusion_matrix(y_true, y_pred, class_names):
   cm = confusion_matrix(y_true, y_pred)
    plt.figure(figsize=(10, 7))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=
   class_names, yticklabels=class_names)
```

```
    plt.title('Matriz de Confusion')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
```

Programa B.2: Función para mostar la matriz de confusión

```
def plot_roc_curve(fpr, tpr, auc_score, nombre):
    plt.figure(figsize=(8, 6))
    plt.plot(fpr, tpr, lw=2, label=f'(AUC = {auc_score:.4f})')
    plt.plot([0, 1], [0, 1], color='red', lw=2, linestyle='--')
    plt.xlabel('Sensibilidad')
    plt.ylabel('Especificidad')
    plt.title('Curva ROC del conjunto de '+nombre)
    plt.legend(loc='lower right')
    plt.show()
```

Programa B.3: Función para mostrar el gráfico de la cuva ROC

Anexo C

Código del análisis de los genes

En este anexo se encuentra el código utilizado para realizar el análisis de influencia de los genes.

```
coefficients = model.coef_[0]
coef_df = pd.DataFrame({
    'gene': data.columns,
    'coefficient': coefficients
})
# Ordenar el DataFrame por el valor absoluto de los coeficientes
coef_df['abs_coefficient'] = coef_df['coefficient'].abs()
coef_df = coef_df.sort_values(by='abs_coefficient', ascending=False)
# Visualizar los 20 genes mas importantes
top_genes = coef_df.head(20)
plt.figure(figsize=(10, 6))
plt.barh(top_genes['gene'], top_genes['coefficient'])
plt.xlabel('Coeficiente')
plt.ylabel('Gen')
plt.title('Top 20 genes que mas influyen en la deteccion de metastasis
    ')
plt.gca().invert_yaxis()
plt.show()
```

Programa C.1: Visualización de los genes más influyentes en la detección de metástasis

```
positive_coef_df = coef_df[coef_df['coefficient'] > 0]
positive_coef_df = positive_coef_df.sort_values(by='coefficient',
    ascending=False)
top_positive_genes = positive_coef_df.head(20)
```

```
plt.figure(figsize=(10, 6))
plt.barh(top_positive_genes['gene'], top_positive_genes['coefficient'
    ])
plt.xlabel('Coeficiente')
plt.ylabel('Gen')
plt.title('Top 20 genes con los coeficientes positivos mas altos en la
    deteccion de metastasis')
plt.gca().invert_yaxis()
plt.show()
```

Programa C.2: Visualización de los genes con coeficientes positivos más influyentes en la detección de metástasis

Bibliografía

[1] Metástasis. URL https://www.irbbarcelona.org/es/reto-metastasis.

[2] El análisis del arn tumoral quizás ayude a encontrar el tratamiento de cáncer más eficaz para cada paciente, 2020. URL https://www.cancer.gov/espanol/noticias/temas-y-relatos-blog/2020/rna-tumoral-medicina-precision-cancer.

[3] Vaske OM. Comparative tumor rna sequencing analysis for difficult-to-treat pediatric and young adult patients with cancer. 2019. doi: 10.1001/jamanetworkopen.2019.13968. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2753519.

[4] Cáncer. URL https://www.who.int/es/news-room/fact-sheets/detail/cancer.

[5] B. Bagiröz, E. Doruk, and O. Yildiz. Machine learning in bioinformatics: Gene expression and microarray studies. ELECTR NETWORK, November 2020. IEEE. ISBN 978-1-7281-8073-1. URL http://dx.doi.org/10.1109/TIPTEKNO50054.2020.9299285.

[6] Joseph M. De Guia, Madhavi Devaraj, and Larry A. Vea. Cancer classification of gene expression data using machine learning models. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, Baguio City, Philippines, November-December 2018. IEEE. ISBN 978-1-5386-7767-4. URL https://ieeexplore.ieee.org/document/8666435.

[7] Fadi Alharbi and Aleksandar Vakanski. Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering*, 10(2):173, February 2023. doi: 10.3390/bioengineering10020173. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9952758/.

[8] Seokjin Haam, Jae-Ho Han, Hyun Woo Lee, and Young Wha Koh. Tumor nonimmune-microenvironment-related gene expression signature predicts brain metastasis in lung adenocarcinoma patients after surgery: A machine learning approach using gene expression

profiling. *Cancers*, 13(17):4468, September 2021. doi: 10.3390/cancers13174468. URL https://doi.org/10.3390/cancers13174468.

[9] Somayah Albaradei, Mahmut Uludag, Maha A. Thafar, Takashi Gojobori, Magbubah Essack, and Xin Gao. Predicting bone metastasis using gene expression-based machine learning models. *Frontiers in Genetics*, 12:771092, November 2021. doi: 10.3389/fgene.2021.771092. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8631472/.

[10] Jinmyung Jung and Sunyong Yoo. Identification of breast cancer metastasis markers from gene expression profiles using machine learning approaches. *Genes*, 14(9):1820, September 2023. doi: 10.3390/genes14091820. URL https://doi.org/10.3390/genes14091820.

[11] Jaeyoon Kim, Minhyeok Lee, and Junhee Seok. Deep learning model with l1 penalty for predicting breast cancer metastasis using gene expression data. *Machine Learning: Science and Technology*, 4(2):025026, June 2023. doi: 10.1088/2632-2153/acd987. URL https://iopscience.iop.org/article/10.1088/2632-2153/acd987.

[12] The cancer genome atlas program. URL https://www.cancer.gov/ccg/research/genome-sequencing/tcga.

[13] Joe W. Chen and Joseph Dhahbi. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci Rep.*, June 2021. doi: 10.1038/s41598-021-92725-8. URL https://doi.org/10.1038%2Fs41598-021-92725-8.

[14] Jiande Wu and Chindo Hicks. Breast cancer type classification using machine learning. *J Pers Med.*, January 2021. doi: 10.3390/jpm11020061. URL https://doi.org/10.3390%2Fjpm11020061.

[15] Suli Liu and Wu Yao. Prediction of lung cancer using gene expression and deep learning with kl divergence gene selection. *BMC Bioinformatics.*, May 2022. doi: 10.1186/s12859-022-04689-9. URL https://doi.org/10.1186%2Fs12859-022-04689-9.

[16] R. URL https://www.r-project.org/.

[17] Bioconductor. URL https://www.bioconductor.org/about/.

[18] Tcgabiolinks. URL https://www.bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html.

[19] Python, . URL https://www.python.org/doc/essays/blurb/.

[20] Imbalanced-learn. URL https://imbalanced-learn.org/stable/.

[21] Scikit-learn. URL https://scikit-learn.org/stable/.

[22] Xgboost, . URL https://xgboost.readthedocs.io/en/stable/.

[23] Tensorflow. URL https://www.tensorflow.org/?hl=es-419.

[24] Pytorch, . URL https://pytorch.org/.

[25] Autogluon. URL https://pytorch.org/.

[26] Optuna. URL https://optuna.org/.

[27] Seaborn. URL https://seaborn.pydata.org/.

[28] Cómo funciona el algoritmo xgboost, . URL https://pro.arcgis.com/es/pro-app/latest/tool-reference/geoai/how-xgboost-works.htm.