I302 - Aprendizaje Automático y Aprendizaje Profundo

Trabajo Práctico 2: Clasificación y Ensemble Learning

Juan Francisco Lebrero

14 de abril de 2025

Ingeniería en Inteligencia Artificial

Diagnóstico de Cáncer de Mama

Resumen

Este trabajo aborda el diagnóstico de cáncer de mama a partir de variables morfológicas y bioquímicas extraídas de imágenes histopatológicas, mediante un enfoque de análisis exploratorio y modelado predictivo. Se identificaron y trataron anomalías estructurales en los datos, como valores atípicos y ausencias, empleando técnicas de imputación por K-Nearest Neighbors y winsorización. Se desarrollaron modelos de regresión logística regularizada sobre conjuntos balanceados y desbalanceados, evaluando distintas estrategias de rebalanceo: *oversampling*, *undersampling*, *SMOTE* y *cost re-weighting*.

En el caso de los datos balanceados, el modelo final alcanzó un recall de 0,8916, una precisión de 0,9250 y un F1-score de 0,9080 al ser evaluado sobre el conjunto de prueba, demostrando una excelente capacidad de generalización. Por su parte, el mejor resultado sobre el conjunto desbalanceado se obtuvo con *cost re-weighting*, logrando un recall de 0,9118 y un AUC-PR de 0,8785. Estos resultados evidencian la solidez de los modelos propuestos en distintos escenarios clínicos.

1. Introducción

El diagnóstico precoz del cáncer de mama resulta fundamental para mejorar las tasas de supervivencia y optimizar los tratamientos clínicos. En este trabajo se estudia un conjunto de datos derivados de imágenes histopatológicas de biopsias mamarias, del cual se extrajeron variables morfológicas y moleculares. El objetivo principal consiste en construir un modelo predictivo capaz de clasificar células como benignas (0) o malignas (1), a partir de 12 variables numéricas (como el tamaño celular, densidad nuclear o tasa mitótica) y 2 variables categóricas (tipo celular y presencia de mutaciones genéticas).

2. Metodología

2.1. Preprocesamiento de Datos

El preprocesamiento se realizó de manera diferenciada sobre dos versiones del conjunto de datos: una con clases balanceadas y otra con clases desbalanceadas. En ambos casos, los datos fueron divididos en conjuntos de entrenamiento (80 %) y validación (20 %), reservando un conjunto independiente para prueba. Se aplicaron técnicas de imputación de valores faltantes y tratamiento de outliers. Además, para el conjunto balanceado se generaron nuevas variables derivadas, mientras que para el conjunto desbalanceado se evaluaron distintas estrategias de re-balanceo durante la etapa de modelado.

2.1.1. Conjunto de Datos Balanceado

El conjunto balanceado presentó una distribución equitativa de la variable objetivo `Diagnosis`, con un 53,79 % de observaciones en la clase 0 y un 46,21 % en la clase 1. Esta condición permitió desarrollar modelos sin aplicar técnicas de re-balanceo.

En la etapa inicial se detectaron valores faltantes en 13 de las 14 variables predictoras. Aproximadamente el 92,26 % de las muestras presentaban al menos un valor ausente. Para su tratamiento, se aplicó imputación mediante el algoritmo de vecinos más cercanos (*K-Nearest Neighbors*), con $k = 8$ y distancia euclidiana como métrica, según el procedimiento detallado en el Apéndice C. El 7,74 % restante de observaciones completas sirvió como base de referencia para esta imputación.

Posteriormente, se identificaron valores atípicos utilizando el método del rango intercuartílico (IQR), como se describe en el

Apéndice D. Los valores extremos fueron corregidos mediante winsorización, ajustándolos al límite correspondiente según el rango definido por los cuartiles. Este procedimiento permitió reducir el impacto de outliers sin eliminar observaciones del conjunto.

A continuación, se analizaron las distribuciones de estas variables numéricas tras el preprocesamiento. La Figura 1 ilustra cómo dichas variables adoptan un comportamiento más simétrico y adecuado para su uso en modelos estadísticos.

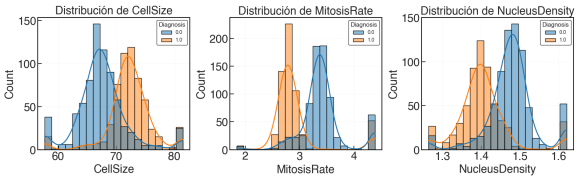

Figura 1: Distribuciones de `CellSize`, `MitosisRate` y `NucleusDensity` tras el preprocesamiento (conjunto balanceado).

Posteriormente, se evaluaron las correlaciones entre las variables numéricas y la variable objetivo `Diagnosis`, utilizando el coeficiente de Pearson. A diferencia del conjunto original, las características mostraron una asociación más clara luego del preprocesamiento, con coeficientes de correlación que oscilaron entre aproximadamente $-0{,}60$ y $0{,}63$. Esta mejora facilitó el análisis posterior y la selección de predictores relevantes.

Las variables categóricas fueron codificadas mediante *One-Hot Encoding*, y las variables numéricas normalizadas en los casos que así lo requerían. Finalmente, se incorporaron variables derivadas mediante técnicas de *feature engineering*, con el objetivo de enriquecer el conjunto de predictores:

- Índice núcleo-citoplasma: (CellSize − CytoplasmSize)/CytoplasmSize,
- Índice de proliferación celular: GrowthFactor × MitosisRate,
- Densidad nuclear: NucleusDensity × ChromatinTexture.

2.1.2. Conjunto de Datos Desbalanceado

En el conjunto desbalanceado, la variable `Diagnosis` presentó un desequilibrio considerable, con un 75 % de observaciones en la clase 0 y un 25 % en la clase 1. Debido a esta asimetría, se evaluaron distintas técnicas de rebalanceo (ver Apéndice A.2) durante la etapa de modelado, pero el preprocesamiento estructural fue equivalente al aplicado sobre los datos balanceados (sección 2.1.1).

A diferencia del conjunto balanceado, en este caso no se aplicó *feature engineering*, con el objetivo de aislar el impacto de las distintas técnicas de re-balanceo sobre el rendimiento del modelo. Para evaluar la calidad de las distribuciones numéricas obtenidas, se generaron visualizaciones análogas a las del conjunto balanceado. La Figura 2 muestra las nuevas distribuciones de las variables `CellSize`, `MitosisRate` y `NucleusDensity` tras el tratamiento de valores faltantes y atípicos.

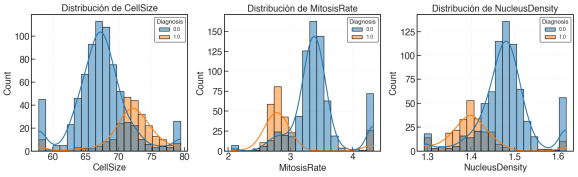

Figura 2: Distribuciones de `CellSize`, `MitosisRate` y `NucleusDensity` tras el preprocesamiento (conjunto desbalanceado).

Las correlaciones observadas entre las variables predictoras y la variable objetivo fueron consistentes con las obtenidas en el conjunto balanceado. Las pequeñas diferencias numéricas no afectaron las asociaciones estructurales principales, confirmando la estabilidad de los patrones estadísticos entre ambos conjuntos.

3. Resultados

A continuación se presentan los principales resultados obtenidos tras la implementación y evaluación de distintas técnicas aplicadas a un modelo de Regresión Logística. Para una descripción detallada del proceso de entrenamiento, ajuste de hiperparámetros, métricas de desempeño y técnicas de rebalanceo, se remite al Apéndice A.1, Apéndice A.2, Apéndice A.3 y Apéndice E, respectivamente.

3.1. Datos Balanceados

El desempeño del modelo fue inicialmente evaluado utilizando el conjunto de validación, lo que permitió obtener una primera estimación de su capacidad predictiva. Posteriormente, se procedió al reentrenamiento del modelo empleando la totalidad del conjunto de desarrollo, aplicando el preprocesamiento desde cero para evitar cualquier tipo de filtración de información y asegurar consistencia con el pipeline original, con el objetivo de aprovechar al máximo la información disponible antes de su evaluación final. Cabe señalar que, si bien el conjunto de validación cumple un rol clave en la comparación y selección entre múltiples modelos o configuraciones, en este caso particular —al tratarse de un único modelo base— se priorizó el análisis sobre el conjunto de prueba como instancia definitiva de evaluación del rendimiento.

En esta etapa se observó una mejora generalizada en las métricas, particularmente en la precisión, que alcanzó el 0,9259. El recall se mantuvo elevado (0,9036) y el F1 score fue de 0,9146, evidenciando un buen equilibrio entre Precision y Recall. La Tabla 3 resume el rendimiento del modelo en ambas fases.

La Figura 3 presenta la curva ROC obtenida sobre el conjunto de prueba, la cual permite visualizar la capacidad discriminativa del modelo frente a distintos umbrales de decisión. Junto con ella, se incorporan la matriz de confusión y la curva Precision-Recall, que complementan el análisis y permiten una evaluación más detallada del desempeño en cada clase. Estos resultados indican que el modelo es capaz de generalizar adecuadamente cuando se entrena sobre datos con distribución equilibrada, sin necesidad de aplicar técnicas adicionales de manejo de desbalanceo.

3.2. Datos Desbalanceados

En esta etapa se evaluó el impacto de distintas técnicas de re-balanceo aplicadas al conjunto de entrenamiento desbalanceado, utilizando validación cruzada para obtener métricas de rendimiento comparables.

3.2.1. Evaluación sobre el Conjunto de Validación

Los resultados obtenidos en la validación muestran un desempeño general sólido para todas las estrategias evaluadas (ver Tabla 4). Sin embargo, dado el carácter clínico del problema —donde minimizar los falsos negativos es prioritario—, el foco se centró en la capacidad de los modelos para identificar correctamente la clase positiva. En este contexto, el recall se estableció como métrica principal, y se observó una mejora considerable en dicha métrica en todos los enfoques con re-balanceo respecto al modelo base.

Entre las técnicas aplicadas, *SMOTE* y *cost reweighting* se destacaron por ofrecer un alto nivel de recuperación de la clase positiva sin sacrificar significativamente otras métricas como la precisión o el F1-score. En particular, *cost reweighting* obtuvo el mayor AUC-PR, lo que refuerza su eficacia en escenarios desbalanceados. Aunque el oversampling tradicional también alcanzó buen recall, su baja precisión sugiere una mayor proporción de falsos positivos, lo cual puede ser perjudicial en contextos clínicos. Por esta razón, *cost reweighting* surge como la opción más equilibrada y recomendable.

3.2.2. Evaluación sobre el Conjunto de Prueba

Si bien el conjunto de prueba debe reservarse idealmente para la evaluación final del modelo seleccionado, en este trabajo se utilizó también con fines exploratorios, con el objetivo de comparar el desempeño general de cada técnica de re-balanceo en un entorno no visto durante el entrenamiento ni la validación. Esto permite observar la robustez de cada enfoque frente a nuevos datos y extraer conclusiones más informadas sobre su aplicabilidad en contextos reales.

Como se observa en la Tabla 5, y tal como se anticipaba a partir de los resultados en el conjunto de validación, la técnica de *cost re-weighting* logró el mejor equilibrio general entre las métricas evaluadas. Por otro lado, en la Figura 4 se observan las curvas ROC y PR de los modelos entrenados con las diferentes técnicas de rebalanceo. Se observa que, además de mantener un recall elevado, obtuvo el mayor AUC-PR (0,8785), lo que refleja una notable capacidad para discriminar correctamente la clase minoritaria en un contexto desbalanceado. A diferencia del *oversampling*, que puede inducir sobreajuste al replicar instancias, el *reweighting* ajusta la función de pérdida sin modificar la distribución original de los datos, preservando así la diversidad del conjunto de entrenamiento.

3.2.3. Elección del modelo final

De acuerdo con el criterio clínico adoptado, se priorizó el recall como métrica principal a la hora de elegir un modelo final. En línea con esta decisión, el modelo basado en *cost re-weighting* resultó ser el más adecuado. El modelo final fue reentrenado sobre el conjunto completo de desarrollo, manteniendo el esquema de preprocesamiento y ajustando el hiperparámetro de regularización mediante validación cruzada con un valor óptimo de $\lambda = 4{,}2813$.

La evaluación final del modelo sobre el conjunto de prueba evidenció un rendimiento consistente, con una accuracy de 0,9265, una precisión de 0,8158, un recall de 0,9118 y un F1-score de 0,8611. Estos resultados reflejan un buen equilibrio entre sensibilidad y precisión, y, lo más relevante, una sólida capacidad para identificar correctamente la clase minoritaria, aspecto crucial en tareas de diagnóstico clínico. La Figura 5 complementa estos resultados mediante la representación gráfica de la matriz de confusión, la curva ROC y la curva precision-recall, ofreciendo una visión más detallada del comportamiento del modelo frente a diferentes umbrales de decisión. Estos resultados reafirman que el modelo seleccionado ofrece un desempeño robusto y equilibrado, con excelente capacidad de generalización frente a datos desbalanceados, sin requerir la generación ni eliminación de observaciones artificiales.

4. Conclusiones

Este trabajo abordó el desafío del diagnóstico automatizado de cáncer de mama mediante un pipeline integral que incluyó análisis exploratorio, limpieza avanzada de datos y entrenamiento de modelos de clasificación. El preprocesamiento —que comprendió imputación con KNN, tratamiento de outliers y generación de variables derivadas— mejoró notablemente la calidad del conjunto de entrenamiento y permitió identificar relaciones más claras entre las variables predictoras y la clase objetivo, lo cual favoreció la construcción de modelos más eficientes y con mayor poder explicativo.

Frente al desbalance de clases, se compararon distintas estrategias, destacándose *cost re-weighting* por su capacidad para mantener un alto nivel de recall sin alterar la distribución original de los datos. Esta técnica resultó ser la más adecuada para el contexto clínico planteado, donde los falsos negativos tienen

un costo elevado. El modelo final, evaluado sobre un conjunto independiente, alcanzó un recall de 0.9118 y un AUC-PR de 0.8785, evidenciando su robustez y utilidad práctica en escenarios reales de diagnóstico médico.

Clasificación de Jugadores de Baloncesto

Resumen

Este trabajo aborda la clasificación del impacto de jugadores de baloncesto profesional, utilizando la métrica `WAR_class` derivada de `war_total`. Se aplicaron técnicas de preprocesamiento sobre un conjunto de datos reales, incluyendo imputación por KNN y winsorización por IQR. Se entrenaron y evaluaron tres modelos supervisados: regresión logística, análisis discriminante lineal (LDA) y Random Forest, empleando validación cruzada y F1-score ponderado como métrica principal.

Los resultados indican que Random Forest alcanzó el mejor desempeño global, con un F1-score de 0,9580 en el conjunto de prueba, mostrando alta precisión y robustez. LDA obtuvo resultados estables (F1 = 0,9020), mientras que la regresión logística mejoró significativamente tras calibración. El modelo propuesto es reproducible, eficiente y adaptable a tareas similares de clasificación multiclase en contextos deportivos.

5. Introducción

Evaluar el impacto individual de los jugadores en deportes de equipo es fundamental para la toma de decisiones técnicas y estratégicas. En el baloncesto profesional, la métrica *Wins Above Replacement* (WAR) permite estimar el aporte global de un jugador respecto a un reemplazo promedio, integrando múltiples aspectos del rendimiento en una sola medida cuantitativa.

Con el fin de facilitar su interpretación y aplicación práctica, este trabajo plantea la clasificación del impacto de los jugadores en tres niveles discretos: negativo, nulo y positivo. Para ello, se desarrolló un pipeline completo de modelado predictivo, que incluye limpieza de datos, imputación de valores inválidos, tratamiento de outliers, y comparación de modelos supervisados: regresión logística, análisis discriminante lineal (LDA) y Random Forest.

El objetivo principal es identificar el modelo con mejor desempeño y capacidad de generalización para esta tarea multiclase, evaluando su precisión, robustez y viabilidad de implementación en contextos reales.

6. Métodos

6.1. Exploración de Datos y Preprocesamiento

Con el objetivo de garantizar una evaluación robusta, se aplicó una partición estratificada sobre el conjunto `dev`, asignando el 80 % de los datos a entrenamiento y el 20 % restante a validación. Esta partición mantuvo la proporción original de clases de la variable objetivo `war_class`, la cual clasifica el impacto de cada jugador en tres niveles: negativo, nulo y positivo.

La variable `war_class` fue construida a partir de la discretización de la métrica continua `war_total`, que posteriormente fue excluida del conjunto de predictores para evitar fuga de información durante el entrenamiento. El análisis exploratorio y las tareas de preprocesamiento fueron realizadas exclusivamente sobre el conjunto de entrenamiento, asegurando la integridad del esquema de validación.

La distribución de clases en el conjunto de entrenamiento mostró un leve desbalance: la clase correspondiente al rendimiento negativo (clase 0) representó el 37 % de las observaciones, mientras que las clases 1 (nulo) y 2 (positivo) concentraron el 33 % y 30 %, respectivamente. Esta ligera asimetría no justificó la necesidad de aplicar técnicas de re-balanceo.

En cuanto a la calidad del conjunto, no se encontraron valores faltantes explícitos ni duplicados. No obstante, se identificaron valores negativos inverosímiles en variables co-

mo `poss` y `mp`, los cuales no tienen sentido semántico en el contexto (por ejemplo, no es posible registrar minutos jugados negativos). Estas observaciones fueron tratadas como ausentes y posteriormente imputadas utilizando el algoritmo de vecinos más cercanos (*K-Nearest Neighbors*), según se detalla en el Apéndice C.

Para mitigar la influencia de valores extremos, se aplicó una técnica de winsorización basada en el rango intercuartílico (IQR), cuyo procedimiento completo puede consultarse en el Apéndice D. Esta técnica permitió reducir el impacto de los outliers sin eliminar observaciones, ajustando los valores extremos a límites definidos por los cuartiles del conjunto.

Finalmente, el análisis de correlación de Pearson confirmó la relevancia de estas variables para la tarea de clasificación: `raptor_total` mostró una correlación fuerte con la clase objetivo ($\rho \approx 0{,}82$), mientras que `poss` y `mp` presentaron correlaciones moderadas ($\rho \approx 0{,}57$). Estos resultados respaldan su inclusión como predictores en el modelado supervisado.

7. Resultados

A continuación se presentan los principales resultados obtenidos tras la implementación y evaluación de distintos modelos predictivos. La descripción completa del proceso de modelado —incluyendo la configuración, el esquema de entrenamiento y las estrategias de evaluación utilizadas para cada algoritmo— se encuentra detallada en el Apéndice B.1 (Modelos utilizados) y en el Apéndice E (Métricas de desempeño).

7.1. Evaluación Inicial sobre el Conjunto de Validación

Una vez definidos y entrenados los modelos sobre el conjunto de entrenamiento, se procedió a evaluar su desempeño utilizando un conjunto de validación estratificada. Esta fase buscó estimar la capacidad predictiva preliminar de cada enfoque, utilizando métricas estándar de clasificación multiclase: *accuracy*, *precision*, *recall* y *F1-score*. Los resultados obtenidos se presentan en la Tabla 7.

Los resultados de esta primera etapa mostraron una clara superioridad del modelo Random Forest, que alcanzó valores cercanos al 98 % en todas las métricas. Este desempeño indica una notable capacidad para capturar relaciones complejas entre las variables predictoras y la clase objetivo. En segundo lugar, se ubicó el modelo LDA, con métricas superiores al 92 %, lo que demuestra su buena adecuación estructural al problema, particularmente bajo los supuestos gaussianos. En contraste, la regresión logística obtuvo resultados considerablemente más bajos, en especial en términos de precisión y F1-score, lo cual sugiere una dificultad para modelar relaciones no lineales dentro del espacio de características.

7.2. Evaluación Final sobre el Conjunto de Prueba

Una vez seleccionados los modelos más competitivos y calibrados sus hiperparámetros, se procedió al reentrenamiento utilizando la totalidad del conjunto de desarrollo (entrenamiento + validación), replicando exactamente el mismo esquema de preprocesamiento aplicado en las fases anteriores. Esta estrategia permitió consolidar el aprendizaje sobre un mayor volumen de datos sin comprometer la validez del conjunto de prueba independiente. Los resultados finales se detallan en la Tabla 6.

Los resultados obtenidos refuerzan las conclusiones preliminares: Random Forest mantuvo su posición como el modelo de mejor rendimiento, con un F1-score de 0.9580 en el conjunto de prueba, muy cercano al valor alcanzado en validación, lo que evidencia su excelente capacidad de generalización y estabilidad. LDA también mostró un comportamien-

to coherente entre ambas fases, con métricas altas y consistentes. Por otro lado, la regresión logística multiclase exhibió una mejora destacada, con un incremento de aproximadamente un 78 % en su F1-score respecto a la validación (de 0,4965 a 0,8836). Este salto sugiere que el modelo fue capaz de beneficiarse significativamente del mayor volumen de datos en el reentrenamiento, y que posiblemente la configuración inicial subestimó su verdadero potencial predictivo.

7.3. Selección del Modelo Final

La elección del modelo a implementar en un entorno productivo no debe basarse únicamente en el desempeño cuantitativo, sino también en criterios de estabilidad, interpretabilidad, eficiencia computacional y adaptabilidad a nuevos datos. Considerando estos factores, el modelo Random Forest emerge como la opción más robusta y eficaz, no sólo por sus métricas superiores, sino también por su baja variabilidad entre validación y prueba. Su arquitectura basada en árboles permite capturar interacciones no lineales sin necesidad de transformaciones previas complejas, lo que facilita su integración en entornos reales.

Además, visualizaciones complementarias —como las matrices de confusión y las curvas ROC y PR generadas para cada modelo— confirmaron la capacidad del Random Forest para minimizar errores de clasificación entre clases adyacentes, un aspecto crítico en tareas donde las decisiones están vinculadas a niveles de rendimiento ordinal. Para un mayor detalle, en la Figura 6 se observa el gráfico conjunto de ROC, matriz de confusión y PR del modelo evaluado sobre el conjunto de prueba.

En conjunto, los resultados validan la elección de Random Forest como modelo final recomendado para la tarea de clasificación de impacto deportivo, resaltando su versatilidad, precisión y capacidad de generalización en contextos multiclase.

8. Conclusiones

Este trabajo abordó la clasificación del impacto de jugadores de baloncesto profesional utilizando la métrica `WAR_class`, construyendo un pipeline completo desde el preprocesamiento hasta la evaluación comparativa de modelos. El análisis inicial permitió detectar y corregir inconsistencias en los datos mediante imputación por KNN y winsorización por IQR, mejorando así la calidad de los insumos para el modelado. Sobre esta base, se entrenaron tres algoritmos supervisados: regresión logística, análisis discriminante lineal (LDA) y Random Forest, priorizando el F1-score ponderado como métrica de referencia y aplicando validación cruzada estratificada para la calibración de hiperparámetros.

Los resultados demostraron una clara superioridad del modelo Random Forest, que logró una alta precisión y robustez tanto en validación como en prueba, adaptándose adecuadamente a la naturaleza multiclase y a la complejidad del problema. Si bien LDA mostró un rendimiento sólido y estable, y la regresión logística mejoró tras su reentrenamiento, Random Forest se destacó por su capacidad para modelar interacciones no lineales, su tolerancia a valores atípicos y su consistencia frente a nuevas observaciones.

Apéndice

A. Diagnóstico de Cáncer de Mama

Este apéndice reúne todas las figuras y tablas incluidas a lo largo del análisis del Diagnóstico de Cáncer de Mama. Cada recurso visual o tabular se encuentra correctamente referenciado en el cuerpo principal, y aquí se presentan organizados para facilitar su consulta y comparación posterior.

A.1. Modelado Predictivo

El modelo utilizado fue regresión logística regularizada para clasificación binaria. La probabilidad de que una observación $\mathbf{x}$ pertenezca a la clase positiva se modela mediante la función sigmoide:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta} \cdot \mathbf{x})}}. \tag{1}$$

La función de pérdida base se define como la entropía cruzada:

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{2}$$

donde $\hat{y}_i = P(y_i = 1 \mid \mathbf{x}_i)$. Para mitigar el sobreajuste, se aplicó una penalización L2:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2. \tag{3}$$

El modelo fue optimizado mediante descenso por gradiente, con una tasa de aprendizaje de 0.01, tolerancia de convergencia de 1×10^{-4} y un máximo de 1000 iteraciones.

A.2. Manejo del Desbalanceo

Dado el desbalance observado entre clases, se aplicaron cuatro estrategias de re-balanceo: *undersampling*, *oversampling*, *SMOTE* y *cost reweighting*. La Tabla 2 presenta la distribución de observaciones resultante para cada técnica.

La técnica de **undersampling** reduce aleatoriamente el número de observaciones de la clase mayoritaria hasta igualarlo con el de la clase minoritaria.

En **oversampling**, se replican aleatoriamente ejemplos de la clase minoritaria hasta igualar las cantidades entre clases.

SMOTE (Synthetic Minority Over-sampling Technique) genera nuevas instancias sintéticas de la clase minoritaria interpolando entre observaciones reales cercanas.

Por su parte, **cost reweighting** modifica la función de pérdida del modelo, penalizando con mayor intensidad los errores sobre la clase minoritaria. La función ajustada se define como:

$$\mathcal{L} = -\sum_{i=1}^{n} w_{y_i} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{4}$$

donde los pesos de clase se fijan como $w_0 = 1$ y $w_1 = \pi_2/\pi_1$, siendo π_1 y π_2 las proporciones de la clase minoritaria y mayoritaria, respectivamente.

Cada una de estas estrategias fue implementada por separado y evaluada en términos de su impacto sobre el rendimiento del modelo, con énfasis en su capacidad para detectar correctamente la clase positiva.

A.3. Ajuste de Hiperparámetros

La validación cruzada se utilizó para ajustar el hiperparámetro de regularización λ. Se exploró un rango logarítmico entre 10^{-6} y 10^{1}. En todos los escenarios se mantuvo la misma configuración de entrenamiento, variando únicamente la estrategia de re-balanceo. Para el caso de datos balanceados, se obtuvo $\lambda = 10{,}0$. Por otro lado, para el conjunto de datos desbalanceados, los valores óptimos obtenidos para cada técnica se resumen en la Tabla 1.

A.4. Figuras

A continuación se presentan las principales visualizaciones utilizadas en el desarrollo del informe. Las mismas se encuentran organizadas por temática: tratamiento de valores atípicos, análisis de distribuciones, y evaluación de modelos.

Evaluación de Modelos

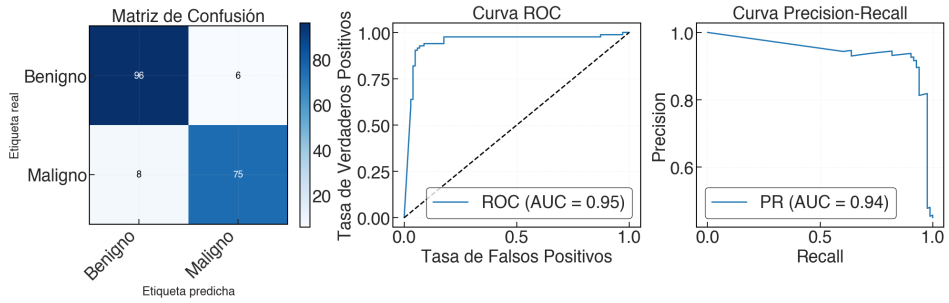

Figura 3: Curva ROC del modelo entrenado sobre datos balanceados y evaluado en el conjunto de prueba.

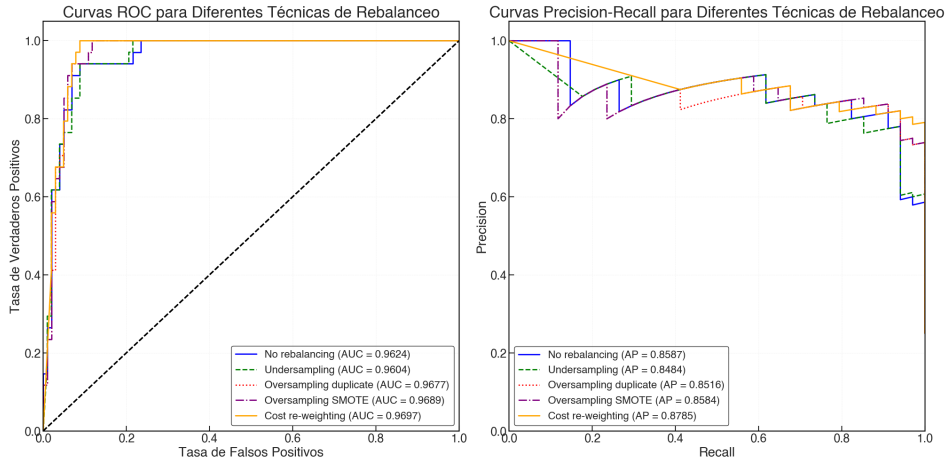

Figura 4: Curvas ROC y PR de los modelos entrenados con distintas técnicas de rebalanceo (conjunto desbalanceado).

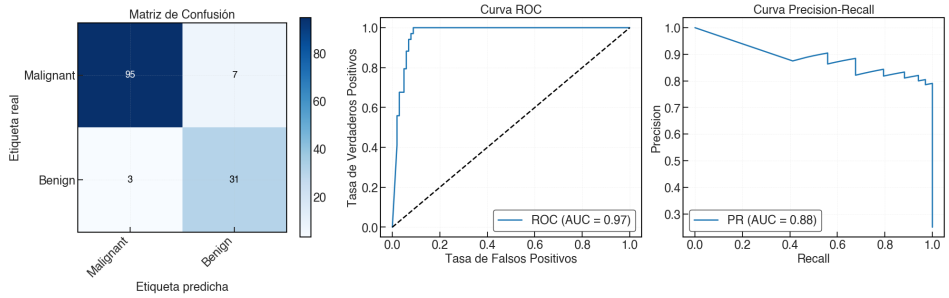

Figura 5: Resumen gráfico del modelo final entrenado con *cost re-weighting*. Incluye matriz de confusión, curva ROC y curva PR.

A.5. Tablas

Cuadro 1: Valores óptimos de λ por técnica de re-balanceo.

Técnica de re-balanceo	**Valor óptimo de λ**
Sin re-balanceo	10
Undersampling	1.27×10^{-5}
Oversampling	0.78
SMOTE	4.28
Cost reweighting	0.34

Cuadro 2: Técnicas de re-balanceo aplicadas y distribución/clasificación resultante.

Técnica	**Clase 0**	**Clase 1**	**Pesos**
Sin re-balanceo	732	243	–
Undersampling	243	243	–
Oversampling	732	732	–
SMOTE	732	732	–
Cost reweighting	732	243	1.0 (0) / 3.0 (1)

Cuadro 3: Desempeño del modelo con datos balanceados en validación y prueba.

Métrica	**Validación**	**Prueba**
Accuracy	0.9066	0.9189
Precision	0.8815	0.9250
Recall	0.8881	0.8916
F1 Score	0.8848	0.9080

Cuadro 4: Resumen de métricas en validación para cada técnica de re-balanceo.

Técnica	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**AUC-PR**
No rebalancing	0.9050	0.8136	0.8000	0.8067	0.9571	0.8468
Undersampling	0.9132	0.8000	0.8667	0.8320	0.9584	0.8468
Oversampling	0.8967	0.7465	0.8833	0.8092	0.9641	0.8676
SMOTE	0.9174	0.8030	0.8833	0.8413	0.9642	0.8654
Cost re-weighting	0.9132	0.7910	0.8833	0.8346	0.9668	0.8819

Cuadro 5: Resumen de métricas en el conjunto de prueba para cada técnica de re-balanceo.

Técnica	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**AUC-PR**
No rebalancing	0.9044	0.8621	0.7353	0.7937	0.9624	0.8587
Undersampling	0.9044	0.8387	0.7647	0.8000	0.9604	0.8484
Oversampling	0.9338	0.8378	0.9118	0.8732	0.9677	0.8516
SMOTE	0.9265	0.8529	0.8529	0.8529	0.9689	0.8584
Cost re-weighting	0.9265	0.8158	0.9118	0.8611	0.9697	0.8785

B. Clasificación de Jugadores de Baloncesto

Este apéndice reúne todas las figuras y tablas incluidas a lo largo del análisis de la Clasificación de Jugadores de Baloncesto. Cada recurso visual o tabular se encuentra correctamente referenciado en el cuerpo principal, y aquí se presentan organizados para facilitar su consulta y comparación posterior.

B.1. Modelado Predictivo

Para abordar la tarea de clasificación, se implementaron distintos enfoques de modelado supervisado, evaluando sus capacidades predictivas sobre el conjunto de datos procesado.

B.1.1. Regresión Logística Multiclase

La regresión logística multiclase utiliza la función softmax para modelar la probabilidad de clase:

$$P(y = k \mid \mathbf{x}) = \frac{e^{\beta_0^{(k)} + \boldsymbol{\beta}^{(k)} \cdot \mathbf{x}}}{\sum_{j=1}^{K} e^{\beta_0^{(j)} + \boldsymbol{\beta}^{(j)} \cdot \mathbf{x}}}. \tag{5}$$

La función de pérdida sin regularización es la entropía cruzada:

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} 1(y_i = k) \log P(y_i = k \mid \mathbf{x}_i), \tag{6}$$

y su versión regularizada con penalización L2 es:

$$\mathcal{L}_{\text{reg}}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{K} \|\boldsymbol{\beta}^{(k)}\|_2^2. \tag{7}$$

La calibración del parámetro de regularización λ se realizó mediante validación cruzada estratificada (5 folds), explorando un rango logarítmico entre 10^{-6} y 10^{1}. Se eligió como métrica de evaluación el F1-score ponderado, y se obtuvo un valor óptimo de $\lambda = 10$. Además, el modelo fue entrenado con los datos normalizados con el fin de garantizar que todas las variables numéricas tuvieran la misma escala, lo cual resulta fundamental para evitar sesgos en el proceso de optimización y asegurar una correcta convergencia del algoritmo de descenso por gradiente.

B.1.2. Análisis Discriminante Lineal (LDA)

El modelo LDA asume que cada clase se distribuye según una normal multivariada con media $\boldsymbol{\mu}_c$ y covarianza común $\boldsymbol{\Sigma}$. Es por esto que el modelo utilizado fue entrenado con los datos normalizados. La función discriminante es:

$$\delta_c(\mathbf{x}) = \gamma_c + \boldsymbol{\beta}_c^{\top} \mathbf{x}, \tag{8}$$

donde

$$\boldsymbol{\beta}_c = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c, \quad \gamma_c = \log \pi_c - \frac{1}{2} \boldsymbol{\mu}_c^{\top} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_c. \tag{9}$$

La clase predicha es la que maximiza $\delta_c(\mathbf{x})$. Para obtener probabilidades predictivas, se aplica una normalización tipo softmax.

B.1.3. Random Forest

El modelo Random Forest se construyó como un conjunto de árboles entrenados sobre subconjuntos *bootstrap*. En cada nodo, se eligió el umbral que maximizó la ganancia de información:

$$\text{IG} = \mathcal{I}(y) - \frac{m_L}{m}\mathcal{I}(y_L) - \frac{m_R}{m}\mathcal{I}(y_R), \tag{10}$$

donde $\mathcal{I}$ representa la entropía. La predicción final del modelo se obtuvo mediante votación mayoritaria:

$$\hat{y} = \text{mode}(h_1(x), h_2(x), \ldots, h_T(x)). \tag{11}$$

Se utilizó una configuración compacta: tres árboles con profundidad máxima 10, al menos dos muestras por nodo y una por hoja. Se seleccionaron atributos al azar (`sqrt` del total) y se usó entropía como criterio de división.

B.2. Figuras

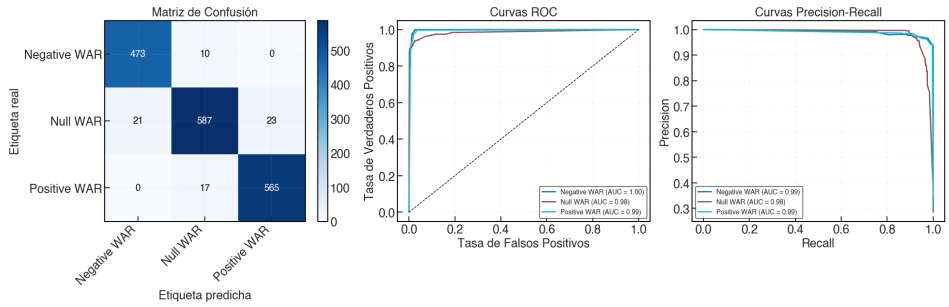

Figura 6: Resumen gráfico de métricas del modelo entrenado con Random Forest y evaluado sobre el conjunto de prueba.

B.3. Tablas

Cuadro 6: Desempeño de modelos sobre el conjunto de prueba `WAR_class_test`

Modelo	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Regresión Logística Multiclase	0.8874	0.8956	0.8874	0.8836
LDA (solver='svd')	0.9051	0.9145	0.9051	0.9020
Random Forest (T=3)	**0.9581**	**0.9581**	**0.9581**	**0.9580**

Cuadro 7: Desempeño de modelos sobre el conjunto de validación

Modelo	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Regresión Logística Multiclase	0.6041	0.4218	0.6041	0.4965
LDA (solver='svd')	0.9232	0.9322	0.9232	0.9215
Random Forest (T=3)	**0.9845**	**0.9847**	**0.9845**	**0.9845**

C. Imputación con KNN

Para el tratamiento de valores faltantes en variables numéricas, se utilizó el algoritmo *K-Nearest Neighbors* (KNN). Este método imputa valores ausentes en función de las observaciones más similares (vecinas), calculadas mediante una métrica de distancia. En este trabajo, se empleó la distancia euclidiana:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{12}$$

donde $\mathbf{x}$ y $\mathbf{y}$ son vectores de características de dos observaciones, e n representa el número de atributos considerados.

D. Winsorización por IQR

Los valores atípicos fueron identificados mediante el método del rango intercuartílico (IQR), una técnica robusta basada en estadísticos de posición. Dado un conjunto de datos, se calcularon los cuartiles Q_1 (percentil 25) y Q_3 (percentil 75), definiendo el IQR como:

$$IQR = Q_3 - Q_1. \tag{13}$$

Los umbrales para considerar un valor como atípico fueron:

$$L_{\text{inf}} = Q_1 - 1{,}5 \times IQR, \quad L_{\text{sup}} = Q_3 + 1{,}5 \times IQR. \tag{14}$$

Los valores fuera de este rango fueron considerados outliers y tratados mediante winsorización, es decir, se truncaron al valor del límite más próximo (L_{inf} o L_{sup}). Esta técnica conserva todas las observaciones, pero limita el impacto de valores extremos sobre el modelo, estabilizando las distribuciones sin alterar la cantidad de datos disponibles.

Este procedimiento fue aplicado a todas las variables numéricas clave antes del entrenamiento de modelos predictivos.

E. Métricas de Desempeño

Para medir el rendimiento, se emplearon las métricas clásicas de clasificación: precisión, recall, F1-score, exactitud, AUC-ROC y AUC-PR. La precisión se define como:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{15}$$

el recall como:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{16}$$

el F1-score como:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{17}$$

y la exactitud como:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{18}$$

Las curvas ROC y PR fueron construidas variando el umbral de decisión. Las áreas bajo dichas curvas (AUC) se utilizaron como indicadores de desempeño global.