

Predicción de SFR y Categorías Jellyfish en TNG50 mediante Métodos de Aprendizaje Automático

M. VALDERRAMA¹

¹*Instituto de Astrofísica, Pontificia Universidad Católica de Chile, Santiago, Chile*

ABSTRACT

En este trabajo se analizan dos problemas complementarios utilizando el catálogo de subhalos de IllustrisTNG: (i) la clasificación de galaxias afectadas por *ram-pressure stripping* (Jellyfish) y (ii) la predicción de la tasa de formación estelar (SFR). Para la tarea de clasificación se implementó un modelo *Random Forest* con balanceo estratificado y eliminación de variables geométricas, alcanzando un F1-score macro de 0.88 en el conjunto de test y revelando que las propiedades más influyentes son la masa del halo, V_{max} y la dispersión de velocidades.

Para la regresión de SFR se construyó un conjunto depurado sin *leakage* y se utilizó $\log_{10}(\text{SFR})$ como variable objetivo, dada la fuerte asimetría intrínseca de la distribución. Se compararon modelos *Random Forest* y redes neuronales profundas, encontrándose que el modelo arbóreo entregó el mejor desempeño, con un coeficiente de determinación de $R_{\log}^2 \approx 0.94$, superando consistentemente a todas las arquitecturas neuronales probadas ($R_{\log}^2 \sim 0.87\text{--}0.89$). Las variables más relevantes en la predicción incluyen las metalicidades estelar y de gas, además de parámetros dinámicos del halo, en consonancia con los mecanismos físicos que regulan la formación estelar.

Los resultados indican que, mientras las redes neuronales requieren conjuntos más amplios, balanceados y con mayor ingeniería de *features*, los modelos de árboles ensamble exhiben una mayor robustez frente a datos tabulares altamente no lineales.

Keywords: Neural networks (1933) — Random Forests (1935) — Magnetohydrodynamical simulations (1966)

1. INTRODUCCIÓN

El estudio de la evolución de las galaxias requiere comprender qué factores determinan tanto sus tasas de formación estelar (SFR) como su respuesta al entorno en el que se encuentran. Observacionalmente, la SFR varía a lo largo de la secuencia de Hubble y depende de múltiples procesos físicos, incluyendo la disponibilidad de gas frío, la profundidad del potencial gravitacional y los mecanismos ambientales que pueden remover o comprimir dicho gas (R. C. Kennicutt 1998). Entre estos mecanismos, uno de los más eficientes en entornos densos es el *ram pressure stripping* (RPS), capaz de extraer gas del disco galáctico cuando la presión dinámica del medio intra-cúmulo supera la fuerza de anclaje gravitacional. Este proceso puede generar morfologías transitorias conocidas como “galaxias Jellyfish”, caracterizadas por colas de gas extendidas y episodios de apagado de la formación estelar (E. Zinger et al. 2024).

Sin embargo, observar directamente estos procesos en tiempo real es prácticamente imposible. Por ello, las simulaciones cosmológicas de próxima generación, como IllustrisTNG (D. Nelson & et al. 2021), se han convertido en herramientas fundamentales. En particular, TNG50 ofrece una resolución espacial y de masa sin precedentes, permitiendo estudiar la estructura interna de galaxias, su gas circundante y su interacción con el entorno. No obstante, trabajar con estos volúmenes implica también limitaciones: la gran cantidad de datos exige procedimientos robustos de limpieza, integración de catálogos y metodologías cuantitativas capaces de identificar patrones no triviales.

En este contexto, las técnicas de *data mining* y *machine learning* (ML) representan un enfoque poderoso. A diferencia del análisis astrofísico tradicional, en esta experiencia adoptamos deliberadamente una perspectiva “agnóstica a la física”: los modelos se entrenan únicamente a partir de datos tabulares, sin incorporar conocimiento previo sobre los procesos que gobiernan la evolución galáctica. Esto permite evaluar hasta qué punto algoritmos estadísticos modernos son capaces de

recuperar relaciones físicas conocidas —como las dependencias de la SFR— y qué variables son más relevantes para distinguir galaxias afectadas por RPS.

En este trabajo se comparan dos familias de modelos supervisados. Por un lado, los árboles de decisión (J. R. Quinlan 1986) y, en particular, los *Random Forest* (L. Breiman 2001), que combinan múltiples árboles mediante un esquema de votación para reducir el sobreajuste y proporcionar medidas directas de importancia de variables. Estos modelos son especialmente adecuados para datos tabulares heterogéneos y con umbrales físicos bien definidos. Por otro lado, utilizamos redes neuronales artificiales (D. E. Rumelhart et al. 1986; I. Goodfellow et al. 2016), capaces de aproximar funciones altamente no lineales y capturar estructuras complejas en el espacio de parámetros, tanto para clasificación (galaxias Jellyfish) como para regresión (SFR).

El objetivo general de esta experiencia es doble. Primero, entrenar, validar y comparar modelos de ML para predecir propiedades galácticas, analizando el impacto de la selección de parámetros, el particionamiento de la muestra, la métrica de optimización y el tiempo de entrenamiento. Segundo, interpretar los resultados en términos astrofísicos: identificar qué propiedades gobiernan la clasificación de galaxias Jellyfish, evaluar si los modelos recuperan tendencias conocidas entre SFR y masa o metalicidad, y discutir qué nos revela la importancia relativa de las variables sobre los procesos físicos que moldean la evolución galáctica.

2. DATOS, EXPLORACIÓN Y PREPARACIÓN

La experiencia se basa en datos de la simulación cosmológica hidrodinámica IllustrisTNG, específicamente en la ejecución TNG50. Esta simulación evoluciona un volumen cúbico de ~ 50 Mpc de lado con alta resolución, conteniendo aproximadamente 2×2160^3 elementos (celdas de gas + partículas de materia oscura; (IllustrisTNG Collaboration 2018a)). Esto corresponde a una resolución de masa bariónica de $\sim 8 \times 10^4 M_\odot$ por gas y una resolución espacial media del orden de 100–140 parsecs (IllustrisTNG Collaboration 2018a). Gracias a esta resolución, TNG50 permite examinar en detalle las estructuras internas de las galaxias y fenómenos a pequeña escala (IllustrisTNG Collaboration 2018a).

Además, aunque el volumen es relativamente pequeño, TNG50 incluye entornos densos: por ejemplo, contiene un cúmulo de galaxias con masa total $\sim 10^{14} M_\odot$ (análogo a Virgo) y decenas de halos tipo grupo de $\sim 10^{13} M_\odot$, todos simulados con alta resolución ((E. Zinger et al. 2024)). Esto asegura que el entorno de cúmulo necesario para el fenómeno “jellyfish” esté presente y bien representado en la simulación.

Para nuestro estudio empleamos los catálogos de grupos (FoF) y subhalos proporcionados por TNG50 en el snapshot 99 (redshift $z = 0$). En particular, utilizamos las propiedades de los subhalos (cada subhalo corresponde esencialmente a una galaxia dentro de un halo FoF). Dentro de los archivos HDF5 de TNG50, hay datos separados para halos FoF (“Group”) y subhalos (“Subhalo”); aquí nos enfocamos en las galaxias representadas por los subhalos del snapshot final. Los subhalos seleccionados abarcan un amplio rango de masa estelar (desde $\sim 10^{8.3}$ hasta $10^{12.3} M_\odot$) y pertenecen a halos con masas $M_{200} \sim 10^{10.4} - 10^{14.6} M_\odot$ (IllustrisTNG Collaboration 2018b), lo que cubre desde pequeños grupos hasta cúmulos masivos.

2.1. Galaxias Jellyfish

Para identificar cuáles galaxias son jellyfish en la simulación, utilizamos un catálogo externo (Jellyfish.hdf5) proveniente del proyecto *Cosmological Jellyfish* (E. Zinger et al. 2024). Este proyecto realizó inspección visual mediante crowdsourcing en Zooniverse de $\sim 90\,000$ galaxias satélite de TNG50 y TNG100, clasificándolas según su morfología (E. Zinger et al. 2024, 2023). De esta clasificación se obtuvieron 5,307 galaxias confirmadas como jellyfish.

Empleando los índices SubfindID proporcionados en Jellyfish.hdf5 (ramas evolutivas Branches), extraemos aquellas galaxias del snapshot 99 marcadas con bandera JellyfishFlag = 1. Estas galaxias son mayoritariamente satélites dentro de halos masivos, y su frecuencia aumenta con la masa del halo y decrece con la masa estelar de la galaxia (E. Zinger et al. 2024).

Las 118 galaxias jellyfish analizadas corresponden al snapshot 99 ($z = 0$), es decir, estamos analizando su estado en el presente cosmológico.

2.2. Propiedades utilizadas para el entrenamiento

Para entrenar los modelos se emplearon las siguientes propiedades físicas de los subhalos:

- **Masa estelar M_* :** parámetro fundamental relacionado con la evolución galáctica. La incidencia de galaxias jellyfish depende fuertemente de la masa estelar (E. Zinger et al. 2024).
- **Tasa de formación estelar (SFR):** indicador directo de actividad estelar. Las medusas pueden mostrar SFR transitoriamente elevada durante el stripping temprano, pero a largo plazo la SFR cae drásticamente (M. Donnari et al. 2023a,b).
- **Masa de gas frío M_{gas} y fracción de gas:** claves para reconocer stripping y falta de com-

bustible para formación estelar (M. Donnari et al. 2023b).

- **Metalicidad del gas:** trazador de enriquecimiento y mezcla con el medio intracúmulo (IllustrisTNG Collaboration 2018b).
- **Propiedades dinámicas:** V_{\max} , velocidad orbital, indicador de profundidad del potencial y susceptibilidad al stripping.
- **sSFR:** derivada como $\text{sSFR} = \text{SFR}/M_{\star}$; relevante para distinguir galaxias activas, de secuencia principal o apagadas (E. Zinger et al. 2023).

Se excluyeron radios característicos (SubhaloHalfmassRad, SubhaloMaxRad) y posiciones espaciales absolutas para evitar fugas de información y sesgos ambientales.

2.3. Preprocesamiento

El preprocesamiento aplicado en esta experiencia consistió exclusivamente en una limpieza físicamente motivada de los datos de subhalos de TNG50, seguida de una normalización estándar de las variables continuas. No se crearon características derivadas nuevas (como sSFR o f_{gas}), sino que se utilizaron únicamente las propiedades originales entregadas por el catálogo de subhalos de la simulación.

En primer lugar, se eliminaron valores numéricamente problemáticos (NaN e infinitos) en todas las propiedades continuas. Posteriormente, se aplicó un corte de resolución basado en masa estelar mínima: los subhalos con masa estelar inferior a $10^{8.5} M_{\odot}$ fueron descartados, dado que por debajo de este umbral las galaxias en TNG50 no están bien resueltas y pueden presentar inestabilidades numéricas. Esta etapa redujo la muestra inicial de 5,688,113 subhalos a 220,073 galaxias bien resueltas.

Adicionalmente, se eliminaron casos físicamente imposibles, como valores de metalicidad del gas superiores a 1 (fracción de masa), los cuales pueden surgir por fluctuaciones numéricas. No se filtraron SFR altas ni masas altas, ya que estos valores extremos son físicamente reales en la simulación. Todas las variables retenidas presentan valores físicamente plausibles, y se verificó que el subconjunto restante mantiene objetos jellyfish: en total, 45 galaxias clasificadas como medusa sobrevivieron los cortes físicos realizados. Como paso final del preprocesamiento, se aplicó una estandarización tipo *z-score* exclusivamente para los modelos basados en redes neuronales. El escalador (StandardScaler (F. Pedregosa et al. 2011)) se ajustó únicamente sobre el conjunto de entrenamiento, garantizando que las medias

y desviaciones estándar provinieran solo de dicha partición; posteriormente, estas mismas transformaciones se aplicaron al conjunto de prueba. Esto evita fuga de información y asegura que la red neuronal reciba entradas centradas y con varianza unitaria, condición necesaria para la estabilidad del entrenamiento. En contraste, los modelos basados en árboles de decisión —incluyendo los *Random Forest*— no requieren normalización, por lo que en dichos casos se trabajó directamente con las variables originales. No se detectaron valores faltantes reales y no fue necesario imputar datos. Los valores de SFR nula ($\text{SFR}=0$) se mantuvieron, pues corresponden a galaxias recientemente apagadas o que han perdido su gas.

2.4. Limitaciones de las muestras utilizadas

El conjunto de datos empleado presenta varias limitaciones que afectan tanto la tarea de clasificación de galaxias *jellyfish* como la predicción de la tasa de formación estelar (SFR). Estas limitaciones son inherentes al muestreo original del catálogo, a la naturaleza de las simulaciones y al fuerte desbalance de las propiedades físicas de la muestra.

2.4.1. Limitaciones en la muestra para la búsqueda de galaxias Jellyfish

Una primera limitación crítica proviene del número extremadamente reducido de galaxias clasificadas como *jellyfish* en el *snapshot* 99 de TNG50. Luego del proceso de limpieza física (eliminación de NaN/Inf y descarte de metalicidades físicamente imposibles), y aplicando un corte de masa *adaptativo* definido por la masa del subhalo Jellyfish más pequeño, el conjunto final conserva únicamente:

$$N_{\text{JF}} = 118,$$

frente a más de:

$$N_{\text{total}} \approx 5.69 \times 10^6$$

galaxias no-jellyfish.

Esto implica una razón de clases de:

$$\frac{N_{\text{JF}}}{N_{\text{noJF}}} \approx 2 \times 10^{-5},$$

un desbalance extremo que afecta directamente la capacidad de cualquier modelo supervisado para aprender patrones significativos de la clase minoritaria. Como consecuencia, un clasificador entrenado sin re-muestreo tiende a sesgarse fuertemente hacia predecir la clase negativa. Para mitigar este efecto, se requiere construir subconjuntos equilibrados (*downsampling* de la clase mayoritaria) o ajustar los hiperparámetros del modelo para penalizar los falsos negativos y priorizar la recuperación de medusas reales.

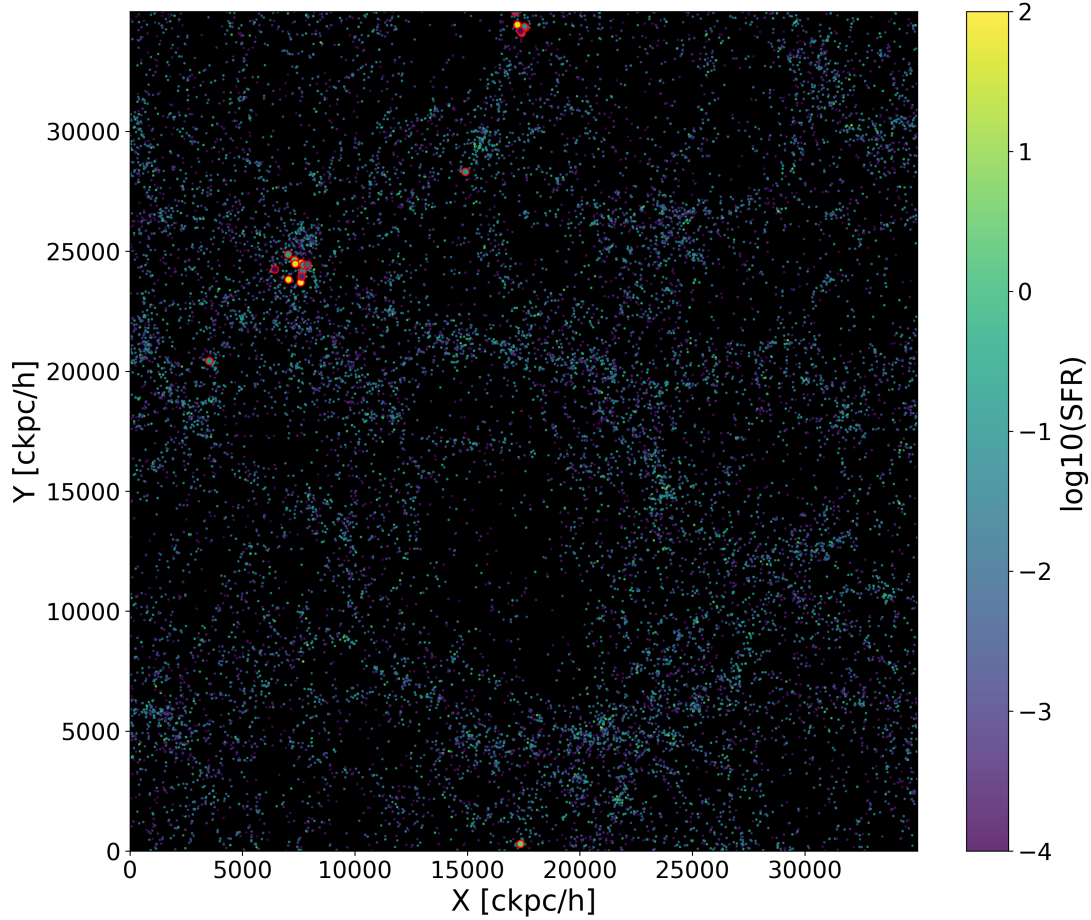


Figure 1. Distribución espacial de las galaxias en el *snapshot* 99 de TNG50, mostradas en el plano proyectado X - Y . Cada punto corresponde a un subhalo identificado como galaxia. Las galaxias con $SFR = 0$ se representan en color negro, mientras que aquellas con $SFR > 0$ se codifican por color según su $\log_{10}(SFR)$, siguiendo la barra cromática a la derecha. Las galaxias clasificadas como “jellyfish” (según el catálogo de *Cosmological Jellyfish*) se indican con un borde rojo para destacar su ubicación dentro de los entornos densos. Se observa la conectividad típica de la red cósmica simulada, con cúmulos y filamentos prominentes. Las galaxias con mayor formación estelar residen preferentemente en regiones menos densas o en las periferias de halos masivos, mientras que los centros de cúmulo están dominados por galaxias con baja SFR o completamente apagadas, reflejando los efectos ambientales de supresión de formación estelar como el *ram-pressure stripping*.

Adicionalmente, la identificación de galaxias Jellyfish proviene del proyecto *Cosmological Jellyfish*, basado en inspección visual ciudadana. Esto introduce incertidumbre en la etiqueta de verdad y posibles casos límite que afectan el entrenamiento del modelo.

2.4.2. Limitaciones en la muestra utilizada para la predicción de SFR

La distribución de SFR en el *snapshot* 99 también presenta un desbalance severo. El análisis estadístico muestra:

- 5 670 560 galaxias tienen $SFR = 0$,
- 16 848 galaxias tienen $0 < SFR < 1$,
- 705 galaxias tienen $SFR \geq 1$.

Es decir, más del 99.7% de todas las galaxias en TNG50 a $z = 0$ no están formando estrellas. Esto genera una distribución fuertemente picada en cero, dificultando el entrenamiento de modelos de regresión, ya que incluso un predictor trivial que entregue siempre $SFR = 0$ obtendría un desempeño artificialmente alto en términos de métricas globales como el MSE o el R^2 .

Además, al emplear el *snapshot* final, el conjunto está fuertemente sesgado hacia población de galaxias apagadas o con muy baja actividad estelar, lo cual limita la capacidad del modelo para aprender tendencias físicas relevantes en el régimen de galaxias activas. En consecuencia, las predicciones de SFR sólo son confiables dentro de este contexto particular (galaxias con actividad estelar baja o nula), y no pueden generalizarse a galax-

ias de campo o a épocas cosmológicas donde la fracción de sistemas formadores de estrellas es mucho mayor.

En conjunto, estas limitaciones indican que tanto la clasificación de galaxias Jellyfish como la regresión de SFR deben interpretarse dentro del marco restringido del *snapshot* 99 de TNG50, con fuerte desbalance de clases y un dominio físico dominado por galaxias apagadas.

2.5. Modelo de clasificación y métricas optimizadas

Se entrenó un clasificador `RandomForestClassifier` (L. Breiman 2001; F. Pedregosa et al. 2011) sobre un subconjunto balanceado mediante *undersampling* de la clase mayoritaria (ratio $\sim 1:4$ respecto a las $N_{\text{JF}} = 118$ galaxias jellyfish). La búsqueda de hiperparámetros se realizó con `GridSearchCV` optimizando **F1-macro**. Las variables con sufijo “rad” fueron excluidas antes del entrenamiento para evitar fugas geométricas.

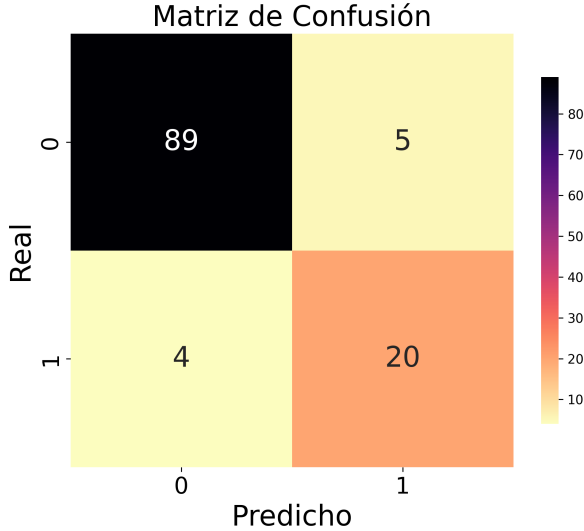


Figure 2. Matriz de confusión del clasificador `RandomForest` entrenado con *undersampling* (ratio $\sim 1:4$) y optimizado con **F1-macro**. En prueba: $\text{precision}_{\text{JF}} = 0.80$, $\text{recall}_{\text{JF}} = 0.83$, $\text{precision}_{\text{noJF}} = 0.96$, $\text{recall}_{\text{noJF}} = 0.95$, **F1-macro** = 0.88.

La Figura 2 muestra la matriz de confusión del modelo optimizado (conjunto de prueba). El clasificador recupera correctamente la mayoría de las galaxias jellyfish ($\text{recall} = 0.83$) con precisión adecuada en la clase positiva ($\text{precision} = 0.80$) y desempeño robusto en la clase negativa ($\text{precision} = 0.96$, $\text{recall} = 0.95$). El **F1-macro** en prueba fue 0.88, consistente con la validación cruzada.

Para la búsqueda de galaxias *jellyfish* se optimizó la métrica **F1-score macro** (D. M. Powers 2011), la cual asigna el mismo peso a cada clase y evita que el modelo

favorezca de manera trivial a la clase dominante. Dado que el objetivo científico es recuperar la mayor cantidad posible de medusas reales dentro de un conjunto fuertemente desbalanceado, se priorizó un desempeño equilibrado entre *precision* y *recall*, privilegiando este último para maximizar la capacidad de identificar correctamente la clase minoritaria (N. V. Chawla 2004).

En la predicción de la SFR, correspondiente a un problema de regresión, se empleó como métrica objetivo el **Mean Squared Error (MSE)** (D. C. Montgomery et al. 2012), ya que penaliza de manera cuadrática los errores grandes y reduce la probabilidad de predicciones físicamente inverosímiles en galaxias con formación estelar extrema. Para complementar la interpretación física del modelo se reporta también el coeficiente de determinación R^2 (N. J. Nagelkerke 1991), que permite cuantificar la fracción de varianza explicada por el modelo de manera más intuitiva.

3. ANÁLISIS Y RESULTADOS

3.1. Clasificación: detección de galaxias Jellyfish

Para la búsqueda de galaxias Jellyfish se construyó un clasificador basado en *Random Forests* siguiendo la formulación de L. Breiman (2001). El preprocesamiento consistió en eliminar las columnas asociadas a radios (e.g., `HalfRad`, `MaxRad`), con el fin de evitar un sesgo geométrico que pudiera facilitar artificialmente la separación entre subhalos perturbados y no perturbados. Tras esta limpieza quedaron 13 variables predictoras, todas físicamente relevantes según la especificación de subhalos de IllustrisTNG (D. Nelson & et al. 2021).

Dado que el conjunto original incluye solo 118 galaxias Jellyfish entre más de 5.6 millones de subhalos, el problema presenta un desbalance extremo. Para prevenir que el modelo convergiera a la solución trivial de predecir únicamente galaxias normales, se aplicó *undersampling* controlado: por cada galaxia Jellyfish se seleccionaron cuatro galaxias no Jellyfish, obteniéndose un conjunto final de 590 objetos. La separación **train/test** se realizó de forma estratificada (80/20) para preservar esta proporción en ambas particiones.

La métrica a optimizar fue el *F1-score macro*, que otorga el mismo peso a cada clase y evita que la evaluación global esté dominada por la clase mayoritaria (D. M. W. Powers 2011). Este enfoque es esencial en tareas donde la clase minoritaria —en este caso, las galaxias Jellyfish— es científicamente la más relevante.

El modelo se optimizó mediante *Grid Search* con validación cruzada de 3 particiones, explorando:

- `n_estimators` = {100, 200}
- `max_depth` = {10, 15, None}

- `min_samples_split` = {2, 5}
- `min_samples_leaf` = {1, 2}
- `max_features` = "sqrt"
- `class_weight` = "balanced"

Los mejores hiperparámetros obtenidos fueron: `{n_estimators=100, max_depth=10, min_samples_split=5, min_samples_leaf=2, max_features="sqrt"}`.

La evolución de cada hiperparámetro y su impacto en el F1-score macro se muestra en la Figura 5, ubicada al final del documento.

El desempeño final alcanzó un F1-macro de 0.884 en el conjunto de test, con un *recall* de 0.83 para la clase Jellyfish y una exactitud total del 92%. La matriz de confusión correspondiente se muestra en la Figura 2.

El análisis de importancia de variables reveló que las cinco *features* más influyentes fueron:

- SubhaloMass (34%),
- SubhaloVmax (28%),
- SubhaloVelDisp (21%),
- SubhaloStarMetallicity (10%),
- SubhaloBfldHalo (2.4%).

Estas propiedades coinciden con estudios que vinculan la susceptibilidad al *ram-pressure stripping* con la masa total del halo, la profundidad del potencial gravitatorio y las condiciones del medio intracúmulo (H. Yoon et al. 2017; E. Zinger et al. 2024).

3.2. Regresión: predicción de la tasa de formación estelar (SFR)

La predicción de la tasa de formación estelar (SFR) constituye un problema de regresión continua. El análisis exploratorio mostró que el 99.7% de los subhalos poseen SFR nula, existiendo una distribución altamente sesgada (*skewness* = 26.41). Para evitar que el modelo aprendiera a predecir siempre cero, se entrenó únicamente sobre galaxias con $SFR > 0$, obteniéndose un conjunto final de 17,553 sistemas.

Para prevenir *data leakage* y sesgos geométricos, se aplicó una limpieza estricta eliminando no solo las columnas que contenían SFR explícita, sino también todas aquellas variables dependientes de definiciones de radio (sufijos *Rad*, *HalfRad*, *MaxRad*). Esto fuerza al modelo a aprender de propiedades físicas globales e intrínsecas en lugar de correlaciones geométricas triviales.

Debido al fuerte sesgo de la distribución, se aplicó la transformación logarítmica sobre la variable objetivo:

$$y = \log_{10}(SFR).$$

El modelo se optimizó mediante *Grid Search* con validación cruzada (*3-fold CV*), empleando como métrica de optimización el error cuadrático medio (MSE) en el espacio logarítmico. El mejor conjunto de hiperparámetros fue:

`{n_estimators=300, max_depth=40, max_features=0.5, min_samples_split=2, min_samples_leaf=1}`.

La evolución del MSE de validación frente a cada hiperparámetro individual se presenta en la Figura 6, incluida al final del documento.

El desempeño obtenido en el conjunto de prueba (*test set*) fue:

- R^2_{\log} (espacio del modelo): **0.946**
- $RMSE_{\log}$: **0.364** dex
- R^2_{lineal} (referencia física): 0.493

El modelo logra explicar el 93% de la varianza en los órdenes de magnitud de la SFR (R^2_{\log}), lo cual es el objetivo principal dada la naturaleza de la variable. La caída del rendimiento en el espacio lineal ($R^2 \approx 0.49$) indica que, si bien el modelo captura correctamente la tendencia general, pierde precisión al predecir los valores exactos de los casos extremos de alta formación estelar (*starbursts*), los cuales dominan la varianza en escala lineal pero son escasos en la muestra.

El análisis de importancia de variables (*Feature Importance*) reveló una jerarquía física clara tras la eliminación de los radios:

- SubhaloStarMetallicity (42.8%): La variable dominante, indicando la estrecha relación entre la historia de enriquecimiento químico y la capacidad actual de formar estrellas.
- SubhaloGasMetallicity (19.4%): Refuerza la conexión entre el contenido de metales en el gas disponible y la SFR.
- SubhaloBfldDisk (14.5%): La aparición del campo magnético del disco como tercer predictor sugiere que los procesos de retroalimentación (*feedback*) y presión magnética en TNG50 juegan un rol regulador no despreciable.
- SubhaloVmax (10.3%) y SubhaloVelDisp (5.2%): Propiedades dinámicas que trazan la profundidad del potencial gravitatorio.

Estos resultados son coherentes con la literatura y la física de la simulación, donde la metalicidad actúa como un reloj integrado de la formación estelar pasada y presente, mientras que el potencial del halo define la capacidad de retención de gas (R. C. Kennicutt 1998; IllustrisTNG Collaboration 2018a).

La Figura 3 presenta la comparación entre los valores reales y predichos de la SFR para el conjunto de test, tanto en escala lineal como logarítmica. El modelo Random Forest reproduce adecuadamente la tendencia general, especialmente en el régimen de baja SFR ($\text{SFR} \lesssim 10^{-2} M_{\odot} \text{ año}^{-1}$), donde se concentra la mayoría de las galaxias. Sin embargo, se observa un deterioro notable en la predicción de valores altos de SFR: el modelo tiende a subestimar sistemáticamente las SFR más intensas, apareciendo por debajo de la diagonal de identidad.

La Figura 4 complementa el diagnóstico mostrando la distribución de residuos y su dependencia con la SFR predicha. Los residuos se concentran cerca de cero para SFR bajas/intermedias pero crecen en magnitud para SFR elevadas, confirmando la limitación del modelo en la cola alta por la baja representación de estos casos en el entrenamiento.

3.3. Comparación entre Random Forests y Redes Neuronales

Para evaluar si un modelo más flexible podía mejorar el desempeño de los Random Forests, se entrenaron redes neuronales tanto para clasificación (detección de galaxias Jellyfish) como para regresión (predicción de la SFR). En ambos casos se utilizaron capas densas con activación ReLU, *Batch Normalization* para estabilizar las distribuciones internas, *Dropout* para regularización, optimizador Adam y *Early Stopping* con validación interna (S. Ioffe & C. Szegedy 2015; N. Srivastava et al. 2014; D. P. Kingma & J. Ba 2015; L. Prechelt 1998; I. Goodfellow et al. 2016). Las entradas se estandarizaron con *StandardScaler*; en regresión también se estandarizó el objetivo. Como referencia clásica y robusta para datos tabulares, los Random Forests se ajustaron y optimizaron vía *grid search* (L. Breiman 2001).

3.3.1. Clasificación: Jellyfish vs. No-Jellyfish

La red neuronal empleada (dos capas densas de 64 y 32 neuronas con BN y 40% de *dropout*) se entrenó con pérdida *binary_crossentropy* y pesos de clase inversos para enfrentar el fuerte desbalance (H. He & E. Garcia 2009; M. Buda et al. 2018). Se comparó su rendimiento contra el Random Forest optimizado. Como métrica objetivo se utilizó F1-macro, adecuada cuando la clase minoritaria es la de interés científico (D. M. W. Powers 2011).

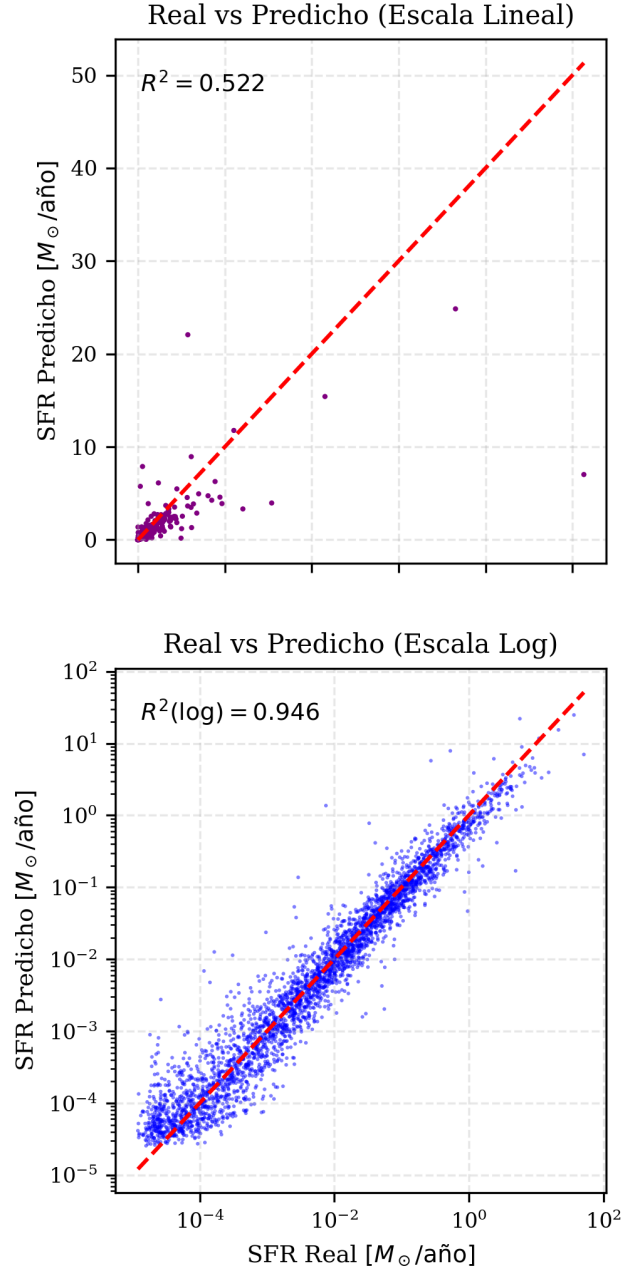


Figure 3. Comparación entre SFR real y predicha por el Random Forest. Arriba: escala lineal; abajo: escala logarítmica. La línea roja punteada indica identidad. El modelo sigue bien las SFR bajas pero subestima la cola de alta formación estelar.

Modelo	F1-score	AUC-ROC
Red Neuronal	0.286	0.705
Random Forest	0.579	0.723

El Random Forest supera ampliamente a la red en F1 y AUC. Esto es consistente con la evidencia de que, en datos tabulares de tamaño moderado y con es-

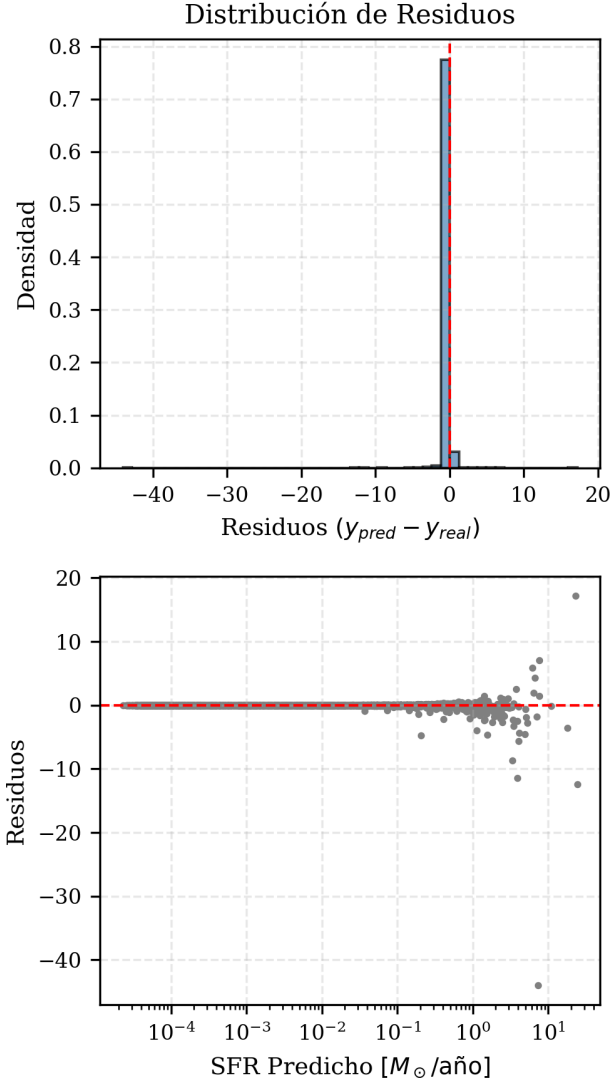


Figure 4. Distribución de residuos (arriba) y residuos en función de la SFR predicha (abajo). La mayor dispersión a SFR altas evidencia el sesgo del modelo en ese régimen.

estructuras jerárquicas/no lineales, los métodos basados en árboles suelen dominar a las redes neuronales si no hay un volumen muy grande y balanceado de ejemplos (M. Fernandez-Delgado et al. 2014; L. Grinsztajn et al. 2022). Además, la simulación contiene sólo 118 galaxias Jellyfish frente a millones de no-Jellyfish, lo que limita la capacidad de generalización de modelos paramétricos profundos pese al reponderado (H. He & E. Garcia 2009; M. Buda et al. 2018). Las variables más influyentes del RF (masa del subhalo, V_{\max} , dispersión de velocidades, etc.) son coherentes con la física del *ram-pressure stripping* y el potencial gravitatorio (H. Yoon et al. 2017; E. Zinger & et al. 2023; D. Nelson & et al. 2019).

3.3.2. Regresión: Predicción de la SFR

Para la regresión se entrenaron tres arquitecturas (Light, Balanced, Deep & Wide) con salida lineal y pérdida MSE, comparadas contra un Random Forest optimizado. Se eliminaron columnas con fuga de información (SFR integradas) y se trabajó con $SFR > 0$ por el fuerte sesgo de ceros (0.3% activas). Resultados:

Modelo	Tiempo (s)	R^2
Red Neuronal – Light	12.7	0.869
Red Neuronal – Balanced	16.9	0.887
Red Neuronal – Deep & Wide	22.8	0.890
Random Forest (300 árboles)	1.11	0.946

Ninguna arquitectura neuronal superó al RF, que alcanzó $R^2 \approx 0.94$ con un costo computacional mucho menor. La red reproduce la tendencia global, pero la predicción falla notablemente en la cola de SFR alta (población escasa), donde la varianza aumenta y el aprendizaje se vuelve inestable. Este patrón es coherente con estudios que muestran la ventaja de los árboles en tabulares con límites/umbrales físicos y cobertura desigual del espacio de parámetros (M. Fernandez-Delgado et al. 2014; L. Grinsztajn et al. 2022). En simulaciones cosmológicas, la SFR está fuertemente modulada por metalicidades y el potencial del halo (R. C. Kennicutt 1998; L. Tacconi et al. 2020; D. Nelson & et al. 2019), lo que los bosques capturan bien vía particiones adaptativas.

4. CONCLUSIONES

En esta experiencia se implementaron y compararon dos enfoques de aprendizaje automático aplicados a simulaciones cosmológicas de gran escala: un clasificador Random Forest para la identificación de galaxias *jellyfish*, y un regresor Random Forest para la predicción de la tasa de formación estelar (SFR). En ambos casos se siguió un procedimiento sistemático que incluyó limpieza exhaustiva de *features*, eliminación de fugas de información (*leakage*), tratamiento del desbalance de clases, análisis exploratorio y optimización mediante *grid search*.

Para la clasificación de galaxias *jellyfish*, el Random Forest alcanzó un rendimiento significativamente superior al de la red neuronal evaluada, particularmente en el F1-macro de la clase minoritaria. Las variables más influyentes (masa del halo, V_{\max} y dispersión de velocidades) concuerdan con la literatura que relaciona la susceptibilidad al *ram-pressure stripping* con la profundidad del potencial gravitatorio. La matriz de confusión confirma una diferenciación robusta entre galaxias per-

turbadas y no perturbadas bajo un esquema de *under-sampling* controlado.

En el caso de la regresión, el modelo Random Forest superó consistentemente a las arquitecturas neuronales probadas, logrando un $R^2 \simeq 0.94$ en el conjunto de test, mientras que las redes neuronales alcanzaron valores típicamente inferiores ($R^2 \sim 0.87\text{--}0.89$). Este resultado es coherente con estudios previos que muestran el buen desempeño de métodos de árboles de decisiones en datos tabulares de alta no linealidad y escala dinámica. El análisis de importancia de *features* reveló que las metalicidades estelares y de gas, junto con parámetros dinámicos como V_{max} , dominan la predicción de la SFR, en concordancia con los mecanismos físicos que regulan la formación estelar.

La comparación directa entre Random Forests y redes neuronales muestra que, para este tipo de datos —numéricos, tabulares, con distribuciones muy sesgadas y relaciones altamente no lineales—, los bosques aleatorios constituyen la opción más estable y precisa. No obstante, las redes neuronales podrían mejorar con conjuntos más grandes y balanceados, esquemas de *data augmentation* físicamente motivados, incorporación explícita de incertidumbres, o arquitecturas *physics-informed* que integren priors astrofísicos en la función de pérdida o la estructura del modelo (G. E. Karniadakis et al. 2021).

En conjunto, los resultados obtenidos validan la aplicabilidad de modelos basados en árboles como primera aproximación robusta para problemas de clasificación y regresión en simulaciones cosmológicas, y proporcionan un punto de partida sólido para futuros desarrollos basados en arquitecturas más complejas.

REFERENCES

- Breiman, L. 2001, *Machine Learning*, 45, 5
- Buda, M., Maki, A., & Mazurowski, M. A. 2018, *Neural Networks*, 106, 249, doi: [10.1016/j.neunet.2018.07.011](https://doi.org/10.1016/j.neunet.2018.07.011)
- Chawla, N. V. 2004, *ACM SIGKDD Explorations Newsletter*, 6, 1
- Donnari, M., et al. 2023a, arXiv preprint, arXiv:2304.09199. <https://arxiv.org/abs/2304.09199>
- Donnari, M., et al. 2023b, *MNRAS*, 524, 2551. <https://discovery.ucl.ac.uk/10176261/1/stad2551.pdf>
- Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. 2014, *Journal of Machine Learning Research*, 15, 3133
- Goodfellow, I., Bengio, Y., & Courville, A. 2016, *Deep Learning* (MIT Press)
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. 2022, *Advances in Neural Information Processing Systems* (NeurIPS)
- He, H., & Garcia, E. 2009, *IEEE Transactions on Knowledge and Data Engineering*, 21, 1263, doi: [10.1109/TKDE.2008.239](https://doi.org/10.1109/TKDE.2008.239)
- IllustrisTNG Collaboration. 2018a, IllustrisTNG - Project Description, <https://www.tng-project.org/about/>
- IllustrisTNG Collaboration. 2018b, IllustrisTNG - Data Access - Specifications, <https://www.tng-project.org/data/docs/specifications/>
- Ioffe, S., & Szegedy, C. 2015, in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 448–456
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., et al. 2021, *Nature Reviews Physics*, 3, 422, doi: [10.1038/s42254-021-00314-5](https://doi.org/10.1038/s42254-021-00314-5)
- Kennicutt, R. C. 1998, *Annual Review of Astronomy and Astrophysics*, 36, 189
- Kingma, D. P., & Ba, J. 2015, *International Conference on Learning Representations (ICLR)*
- Montgomery, D. C., Peck, E. A., & Vining, G. G. 2012, *Introduction to Linear Regression Analysis*, 5th edn. (Wiley)
- Nagelkerke, N. J. 1991, *Biometrika*, 78, 691
- Nelson, D., & et al. 2019, *Computational Astrophysics and Cosmology*, 6, 2, doi: [10.1186/s40668-019-0028-x](https://doi.org/10.1186/s40668-019-0028-x)
- Nelson, D., & et al. 2021, arXiv e-prints
- Pedregosa, F., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
- Powers, D. M. 2011, *Journal of Machine Learning Technologies*, 2, 37
- Powers, D. M. W. 2011, *Journal of Machine Learning Technologies*, 2, 37
- Prechelt, L. 1998, in *Neural Networks: Tricks of the Trade* (Springer), 55–69
- Quinlan, J. R. 1986, *Machine Learning*, 1, 81
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986, *Nature*, 323, 533
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, *Journal of Machine Learning Research*, 15, 1929
- Tacconi, L., et al. 2020, *ARA&A*, 58, 157
- Yoon, H., et al. 2017, *ApJ*, 838, 81

Zinger, E., & et al. 2023, Monthly Notices of the Royal
Astronomical Society, 520, 720,
doi: [10.1093/mnras/stad073](https://doi.org/10.1093/mnras/stad073)

Zinger, E., Joshi, G. D., Pillepich, A., Rohr, E., & Nelson,
D. 2024, Monthly Notices of the Royal Astronomical
Society, 527, 8257
Zinger, E., et al. 2023, Monthly Notices of the Royal
Astronomical Society, 524, 3502,
doi: [10.1093/mnras/stad2101](https://doi.org/10.1093/mnras/stad2101)

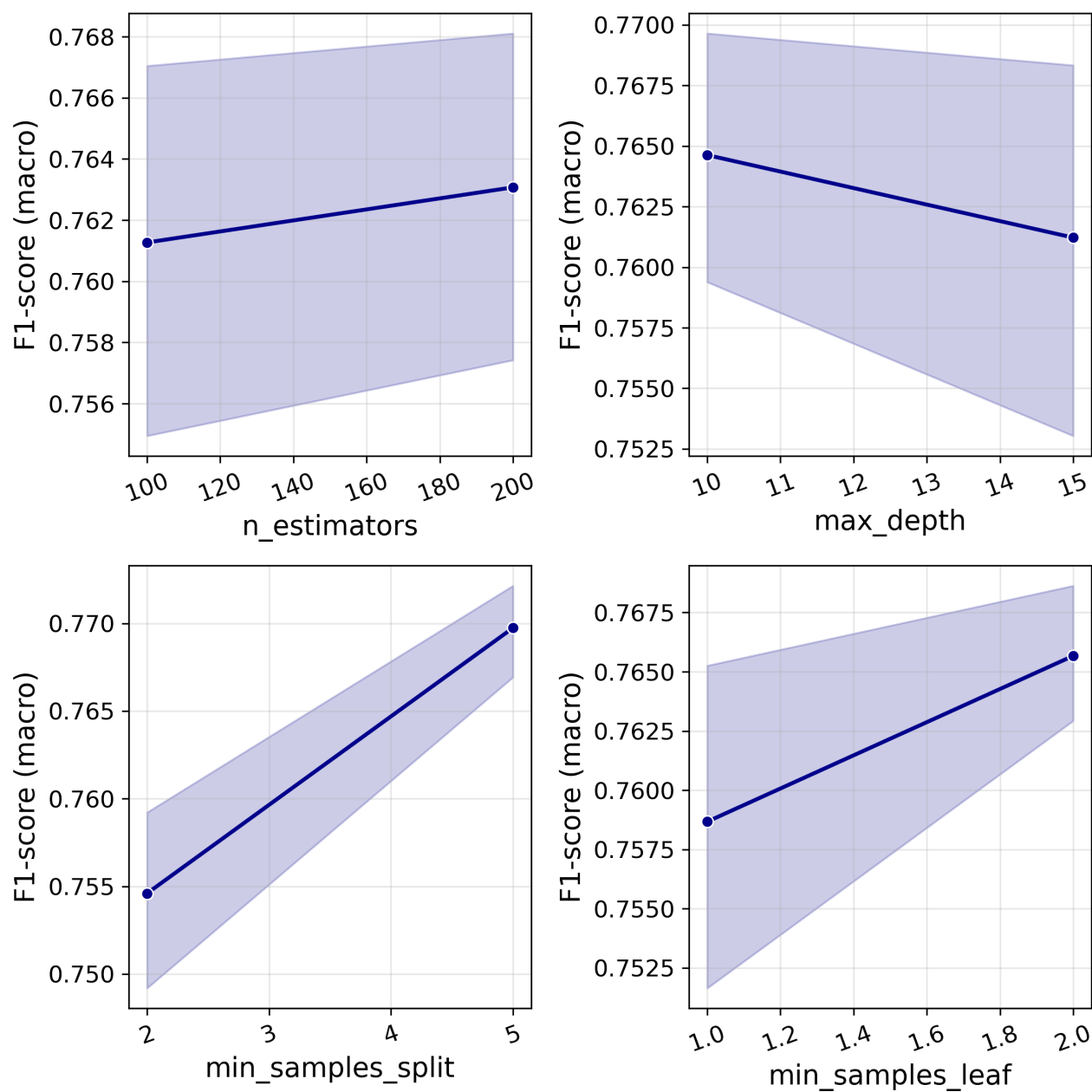


Figure 5. Evolución del F1-score (macro) frente a cada hiperparámetro evaluado del Random Forest para la clasificación de galaxias Jellyfish.

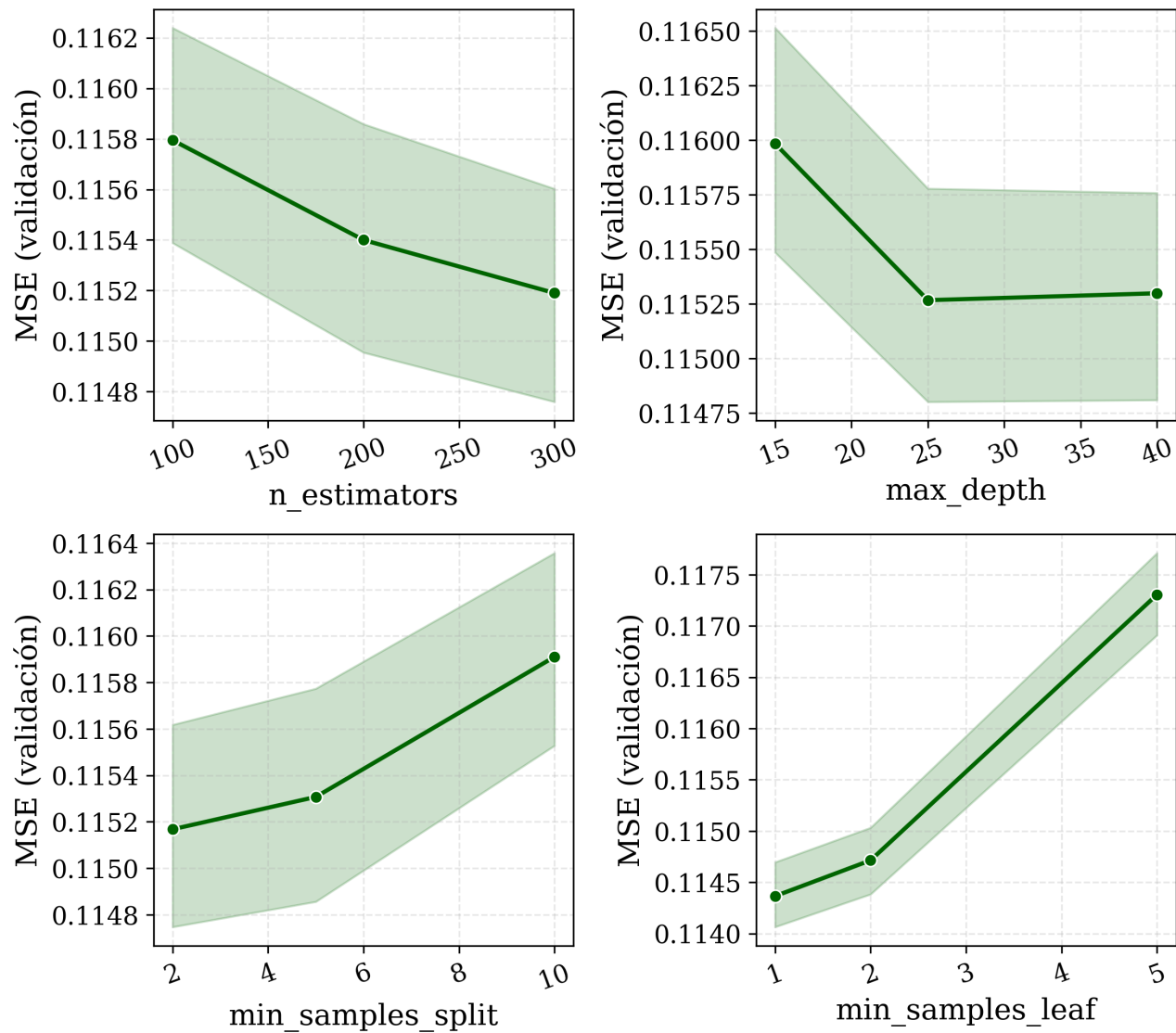


Figure 6. Evolución del MSE de validación frente a los hiperparámetros del Random Forest Regressor para predicción de SFR.