



U N I V E R S I D A D
Panamericana

Minería de Datos: Clasificación Binaria de Estrellas y Galaxias

Juan Pablo de Alba Tamayo

Junio 2025

Índice

1. Introducción	7
2. Definición del Proyecto	7
2.1. Objetivo	7
2.2. Expectativas	7
2.2.1. Métricas de Evaluación	8
3. Exploración de Datos	8
3.1. Descripción de Datos	8
3.1.1. Identificadores y Metadatos del Survey	9
3.1.2. Posición y Movimiento	9
3.1.3. Magnitudes Fotométricas	9
3.1.4. Parámetros Morfológicos	9
3.1.5. Parámetros de Stokes	9
3.1.6. Revisión de Calidad de Datos	10
3.1.7. Estadísticas Descriptivas	11
3.2. Distribución de Datos	12
3.2.1. Balance de Clases	12
3.2.2. Análisis de Distribuciones por Variable	13
3.2.3. Identificación de Valores Atípicos	15
3.3. Correlación de Datos	16
3.3.1. Correlación con la Variable Objetivo	16
3.3.2. Matriz de Correlación Completa	17
3.3.3. Exportación para Análisis Detallado	17
4. Preparación de Datos	18
4.1. Limpieza	18
4.1.1. Verificación de Valores Faltantes y Duplicados	18
4.1.2. Eliminación de Variables No Informativas	18
4.1.3. Codificación de la Variable Objetivo	19
4.2. Selección de Variables	19
4.2.1. Criterios de Selección	19
4.2.2. Variables Seleccionadas	19
4.2.3. Validación Estadística de la Selección	20
4.3. Preprocesado (Logaritmo, RobustScaler)	21
4.3.1. Arquitectura del Pipeline de Preprocesado	21
4.3.2. Transformación	22
4.3.3. Selección de RobustScaler	23
4.3.4. Estrategia de Imputación	24
4.3.5. Optimización de Rendimiento	24
5. Modelo de Machine Learning	25
5.1. Regresión Logística	25
6. Resultados	27
6.1. Comparación de Técnicas de Validación	27
6.2. Optimización de Hiperparámetros	27
6.3. Implementación del Modelo Final	28
6.4. Evaluación en Dataset de Test	28
6.5. Índice Kappa de Cohen	29
6.6. Análisis de la Curva ROC	30

7. Conclusión	31
7.1. Conclusiones del Proyecto	31
7.2. Problemas Enfrentados y Soluciones	31
Bibliografía	33

Índice de figuras

1.	Información general del dataset obtenida con <code>df.info()</code> . Se confirma la estructura completa con 4,000,000 entradas, 51 columnas sin valores nulos, y un uso de memoria de 1.5+ GB. Los tipos de datos incluyen enteros (<code>int64</code>), flotantes (<code>float64</code>) y objetos (<code>object</code>) para la variable objetivo.	11
2.	Estadísticas descriptivas - Ejemplo: Variables de identificación, posición y magnitudes PSF. Se observa que todas las variables tienen el conteo completo de 4,000,000 observaciones, confirmando la ausencia de valores faltantes.	12
3.	Distribución de las clases en el dataset. Se observa que las variables tienen el conteo completo de 4,000,000 observaciones.	12
4.	Histogramas con curvas de densidad KDE de variables con distribuciones relativamente normales. Las curvas de densidad (generadas con <code>sns.histplot</code> y <code>kde=True</code>) proporcionan una representación suave de la distribución, facilitando la identificación de patrones y modas. Se muestran ejemplos representativos de las 52 columnas del dataset.	13
5.	Histogramas con curvas de densidad KDE de variables con distribuciones altamente sesgadas. Las curvas de densidad revelan claramente la concentración de valores cerca de cero y las colas extremas. La estimación por kernel (KDE) ayuda a visualizar mejor la forma de estas distribuciones asimétricas. Debido a la cantidad de columnas (52), se muestran ejemplos representativos.	14
6.	Boxplots de magnitudes fotométricas mostrando la presencia de valores atípicos. Estos outliers representan objetos astronómicos reales (muy brillantes o muy débiles) y contienen información valiosa para la clasificación.	15
7.	Boxplots de parámetros de Stokes y variables de movimiento. Se observa la concentración extrema de valores cerca de cero y la presencia de outliers exagerados, especialmente en los parámetros <code>q_*</code> y <code>u_*</code>	15
8.	Boxplots de parámetros morfológicos (relaciones de ejes y algunos histogramas adicionales). Las variables <code>expAB_*</code> muestran distribuciones más controladas, mientras que otras variables presentan comportamientos diversos. Debido a las 52 columnas del dataset, se presentan ejemplos representativos.	15
9.	Correlación de todas las variables con la variable objetivo (<code>type_numeric</code>). Las variables con correlaciones más altas (en valor absoluto) son las más discriminativas para la clasificación.	16
10.	Matriz de correlación entre todas las variables del dataset. Los colores más intensos indican correlaciones más fuertes (positivas en azul, negativas en amarillo). Se observan bloques de alta correlación entre variables del mismo tipo (e.g., magnitudes en diferentes filtros, radios en diferentes bandas).	17
11.	Estadísticos chi-cuadrado para las variables seleccionadas. Valores más altos indican mayor asociación con la variable objetivo. Todas las variables muestran estadísticos muy elevados, confirmando su relevancia discriminativa.	20
12.	P-valores asociados a la prueba chi-cuadrado para cada variable. Todos los p-valores son prácticamente cero, rechazando la hipótesis nula de independencia.	21
13.	Diagrama del pipeline de preprocesado implementado. Se muestra la arquitectura del ColumnTransformer con dos ramas de procesamiento: una para variables que requieren transformación logarítmica y otra para variables estándar.	22
14.	Ejemplo de la efectividad de la transformación logarítmica en variables con distribuciones altamente sesgadas. Se muestra la comparación entre las distribuciones originales (izquierda) y después de aplicar <code>log1p</code> (derecha) para <code>modelFlux_i</code> y <code>modelFlux_z</code> . La transformación convierte distribuciones extremadamente sesgadas en distribuciones más simétricas y manejables para los algoritmos de machine learning.	23
15.	Pipeline completo para Regresión Logística, integrando el preprocesado especializado para datos astronómicos con el clasificador optimizado.	25
16.	Comparación entre <code>train_test_split</code> (80 %-20 %) y validación cruzada de 5 folds. Ambas técnicas muestran resultados consistentes, confirmando la estabilidad del modelo.	27

17.	Matriz de confusión del modelo final en el dataset de test. Se muestran las predicciones para 1 millón de objetos astronómicos clasificados como STAR (0) o GALAXY (1).	29
18.	Resultado del índice kappa de Cohen para el modelo final. El valor obtenido demuestra una concordancia casi perfecta entre las predicciones del modelo y las etiquetas verdaderas.	29
19.	Curva ROC del modelo final mostrando un AUC de 0.99. La curva se aproxima al punto óptimo (0,1), indicando un rendimiento excelente. La línea punteada representa el rendimiento aleatorio ($AUC = 0.50$).	30

1. Introducción

La astronomía moderna se enfrenta al desafío de procesar y clasificar enormes volúmenes de datos provenientes de observaciones astronómicas. Con el uso de telescopios de gran tamaño y estudios de cielo profundo, la cantidad de objetos celestes detectados ha crecido exponencialmente, haciendo impracticable la clasificación manual de cada uno de estos objetos. En este contexto, la diferenciación entre estrellas y galaxias representa uno de los problemas fundamentales de clasificación en astronomía, ya que estas dos clases de objetos constituyen la mayoría de las fuentes puntuales detectadas en los estudios fotométricos.

Este proyecto tiene como objetivo desarrollar un sistema de clasificación automática para distinguir entre estrellas y galaxias utilizando técnicas de aprendizaje automático aplicadas a datos fotométricos. Se implementará un modelo de Regresión Logística optimizado que aprovecha las características discriminativas más relevantes de los objetos astronómicos. El análisis incluirá la exploración exhaustiva de los datos, el preprocesamiento especializado para datos astronómicos, la selección de características óptimas y la optimización de hiperparámetros para maximizar la eficacia del modelo. La implementación completa del proyecto se encuentra disponible en un notebook interactivo [3].

2. Definición del Proyecto

2.1. Objetivo

El objetivo principal de este proyecto es desarrollar un sistema de clasificación automática que permita distinguir eficientemente entre estrellas y galaxias utilizando técnicas de aprendizaje automático. Específicamente, se busca:

- Implementar un modelo de Regresión Logística optimizado para la clasificación binaria de estrellas y galaxias.
- Realizar un análisis exhaustivo de los datos fotométricos disponibles, incluyendo la exploración de patrones, distribuciones y correlaciones entre variables.
- Aplicar técnicas de preprocesamiento de datos especializadas, como transformaciones logarítmicas y escalado robusto, optimizadas para características astronómicas.
- Optimizar los hiperparámetros del modelo mediante técnicas de búsqueda aleatoria (Randomized-Search) para maximizar su efectividad.
- Desarrollar un pipeline robusto que proporcione excelente capacidad de generalización y precisión superior al 90 % en la clasificación de objetos astronómicos.

2.2. Expectativas

Se espera que al finalizar este proyecto se haya logrado:

- Alcanzar una precisión de clasificación superior al 90 % en el conjunto de datos de prueba, aprovechando la naturaleza bien diferenciada de las características entre estrellas y galaxias.
- Demostrar que la Regresión Logística, con un preprocesamiento adecuado, puede proporcionar un rendimiento excelente y una capacidad de generalización superior para datos astronómicos.
- Identificar las variables fotométricas más discriminativas para la clasificación, contribuyendo al conocimiento sobre qué características observacionales son más útiles para esta tarea.
- Establecer un pipeline de procesamiento de datos robusto y eficiente que pueda ser aplicado a futuros conjuntos de datos astronómicos similares.

2.2.1. Métricas de Evaluación

Para evaluar el rendimiento del modelo de clasificación, se utilizarán las siguientes métricas estándar:

Exactitud (Accuracy): Representa la proporción de predicciones correctas sobre el total de predicciones realizadas. Es la métrica más intuitiva pero puede ser engañosa en datasets desbalanceados [6].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precisión (Precision): Mide la proporción de verdaderos positivos entre todas las predicciones positivas. Responde a la pregunta: "De todos los objetos que el modelo clasificó como galaxias, ¿qué porcentaje realmente son galaxias?" [9]

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Sensibilidad o Recall: Mide la proporción de verdaderos positivos que fueron correctamente identificados. Responde a la pregunta: "De todas las galaxias reales en el dataset, ¿qué porcentaje fue correctamente identificado por el modelo?" [7]

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F1-Score: Es la media armónica entre precisión y recall, proporcionando un balance entre ambas métricas. Es especialmente útil cuando se busca un equilibrio entre no perder objetos importantes (alto recall) y mantener predicciones confiables (alta precisión) [8].

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Área Bajo la Curva ROC (AUC-ROC): Mide la capacidad del modelo para distinguir entre las dos clases a través de todos los posibles umbrales de clasificación. Un valor de 0.5 indica un rendimiento aleatorio, mientras que 1.0 representa una clasificación perfecta [10].

Donde: TP = Verdaderos Positivos, TN = Verdaderos Negativos, FP = Falsos Positivos, FN = Falsos Negativos.

3. Exploración de Datos

3.1. Descripción de Datos

Los datos utilizados en este proyecto provienen del Sloan Digital Sky Survey (SDSS), uno de los estudios astronómicos más comprehensivos jamás realizados [2]. El conjunto de datos se divide en dos partes principales:

- **Dataset de Entrenamiento (train.csv):** Contiene aproximadamente 4 millones de observaciones, divididas equitativamente entre 2 millones de estrellas y 2 millones de galaxias. Este conjunto se utiliza para entrenar y validar los modelos de clasificación.
- **Dataset de Prueba (test.csv):** Contiene 1 millón de observaciones adicionales que se utilizan exclusivamente para evaluar el rendimiento final de los modelos entrenados.

Cada observación en el dataset representa un objeto astronómico y contiene 50 variables que describen diferentes aspectos de sus propiedades observacionales. Estas variables se pueden agrupar en las siguientes categorías:

3.1.1. Identificadores y Metadatos del Survey

- **objID:** Identificador único del SDSS compuesto por varios componentes técnicos del survey
- **run:** Número de la secuencia de observación específica
- **camcol:** Columna de la cámara utilizada (el SDSS tiene 6 columnas de CCD)
- **field:** Número del campo observado dentro de la secuencia
- **type:** Clasificación del tipo de objeto (estrella, galaxia) - esta es nuestra variable objetivo

3.1.2. Posición y Movimiento

- **rowv, colv:** Componentes de velocidad del objeto en grados por día, que pueden indicar movimiento propio (especialmente relevante para estrellas cercanas)
- **ra, dec:** Ascensión recta y declinación, que son las coordenadas estándar para localizar objetos en el cielo (equivalente a longitud y latitud terrestres)
- **b, l:** Latitud y longitud galácticas, que indican la posición del objeto relativa al plano de nuestra galaxia

3.1.3. Magnitudes Fotométricas

Las magnitudes miden el brillo aparente de los objetos en diferentes filtros de color. El sistema fotométrico del SDSS utiliza cinco filtros estándar:

- **psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_z:** Magnitudes PSF (Point Spread Function) en los filtros ultravioleta, verde, rojo, infrarrojo cercano e infrarrojo, respectivamente. Estas magnitudes asumen que el objeto es puntual.
- **u, g, r, i, z:** Alias abreviados para las magnitudes modelo en cada filtro, que representan el mejor ajuste entre modelos exponenciales y de Vaucouleurs.
- **modelFlux_u, modelFlux_g, modelFlux_r, modelFlux_i, modelFlux_z:** Flujos correspondientes a las magnitudes modelo, medidos en nanomaggies (unidad de flujo astronómico).

3.1.4. Parámetros Morfológicos

Estas variables describen la forma y el tamaño aparente de los objetos:

- **petroRad_u, petroRad_g, petroRad_r, petroRad_i, petroRad_z:** Radio petrosiano en cada filtro, que mide el tamaño característico del objeto en segundos de arco.
- **expRad_u, expRad_g, expRad_r, expRad_i, expRad_z:** Radio de ajuste exponencial, también conocido como radio efectivo o de media luz, que indica el tamaño donde se concentra la mitad de la luz del objeto.
- **expAB_u, expAB_g, expAB_r, expAB_i, expAB_z:** Relación de ejes del ajuste exponencial (b/a), que indica qué tan alargado o circular es el objeto.

3.1.5. Parámetros de Stokes

- **q_u, q_g, q_r, q_i, q_z:** Parámetros de Stokes Q en cada filtro, relacionados con la polarización lineal de la luz.
- **u_u, u_g, u_r, u_i, u_z:** Parámetros de Stokes U en cada filtro, también relacionados con la polarización.

Importancia para la Clasificación: Las estrellas, al ser objetos puntuales y relativamente cercanos, tienden a tener radios pequeños y magnitudes PSF bien definidas. Las galaxias, siendo objetos extensos y distantes, muestran estructura morfológica compleja, radios mayores y diferencias significativas entre magnitudes PSF y modelo. Estas diferencias fundamentales en las propiedades observacionales son las que permiten a los algoritmos de machine learning distinguir eficazmente entre ambas clases de objetos.

Antes de proceder con el análisis y modelado, se realizó una evaluación exhaustiva de la calidad y estructura de los datos. Afortunadamente, el dataset del SDSS se encontraba en excelentes condiciones, lo que facilitó significativamente el proceso de análisis.

3.1.6. Revisión de Calidad de Datos

Se llevó a cabo una inspección sistemática para identificar posibles problemas en los datos:

- **Valores Faltantes (NaN):** Se verificó la presencia de valores nulos en todas las variables. El análisis reveló que el dataset no contiene valores faltantes, lo cual es característico de la alta calidad del procesamiento de datos del SDSS.
- **Valores Nulos:** Se confirmó la ausencia de valores nulos en todas las columnas, indicando un dataset completo y consistente.
- **Registros Duplicados:** Se realizó una búsqueda de observaciones duplicadas basada en el identificador único (objID). No se encontraron registros duplicados, confirmando la integridad del dataset.
- **Estructura del Dataset:** El análisis confirmó que el dataset contiene exactamente 51 columnas (variables) como se esperaba, incluyendo todas las características fotométricas, morfológicas y de posición necesarias para la clasificación.

```

RangeIndex: 4000000 entries, 0 to 3999999
Data columns (total 51 columns):
#   column      Dtype
---  ---
0   objID       int64
1   run         int64
2   camcol      int64
3   field       int64
4   type        object
5   rowv        float64
6   colv        float64
7   u           float64
8   g           float64
9   r           float64
10  i           float64
11  z           float64
12  psfMag_u    float64
13  psfMag_g    float64
14  psfMag_r    float64
15  psfMag_i    float64
16  psfMag_z    float64
17  modelFlux_u float64
18  modelFlux_g float64
19  modelFlux_r float64
20  modelFlux_i float64
21  modelFlux_z float64
22  petroRad_u  float64
23  petroRad_g  float64
24  petroRad_r  float64
25  petroRad_i  float64
26  petroRad_z  float64
27  expRad_u    float64
28  expRad_g    float64
29  expRad_r    float64
30  expRad_i    float64
31  expRad_z    float64
32  q_u         float64
33  q_g         float64
34  q_r         float64
35  q_i         float64
36  q_z         float64
37  u_u         float64
38  u_g         float64
39  u_r         float64
40  u_i         float64
41  u_z         float64
42  expAB_u     float64
43  expAB_g     float64
44  expAB_r     float64
45  expAB_i     float64
46  expAB_z     float64
47  ra          float64
48  dec         float64
49  b           float64
50  l           float64
dtypes: float64(46), int64(4), object(1)
memory usage: 1.5+ GB

```

Figura 1: Información general del dataset obtenida con `df.info()`. Se confirma la estructura completa con 4,000,000 entradas, 51 columnas sin valores nulos, y un uso de memoria de 1.5+ GB. Los tipos de datos incluyen enteros (int64), flotantes (float64) y objetos (object) para la variable objetivo.

3.1.7. Estadísticas Descriptivas

Se aplicó la función `df.describe()` para obtener un resumen estadístico completo de todas las variables numéricas del dataset. Este análisis proporcionó información valiosa sobre:

- Medidas de tendencia central (media, mediana)
- Medidas de dispersión (desviación estándar, rango intercuartil)
- Valores mínimos y máximos para cada variable
- Distribución de percentiles (25 %, 50 %, 75 %)

Las estadísticas descriptivas revelaron rangos de valores consistentes con las expectativas astronómicas y confirmaron la ausencia de valores anómalos evidentes que pudieran indicar errores de medición o procesamiento.

Las siguientes figuras muestran el resumen estadístico completo de todas las variables del dataset:

	objid	run	cancel	field	row	colv	u	g	r	i	z	psfmag_u	psfmag_g	psfmag_r	psfmag_i	psfmag_z	modelflux_u	modelflux_g	modelflux_r	modelflux_i	modelflux_z
count	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06	4.000000e+06
mean	1.237648e+18	5.639229e+02	3.491990e+00	3.598391e+02	-1.229906e+00	-1.229669e+00	2.326655e+01	2.196408e+01	2.075499e+01	2.026221e+01	1.993484e+01	2.331607e+01	2.217918e+01	2.103692e+01	2.051476e+01	2.013707e+01	4.850849e+00	1.935893e+01	3.801614e+01	5.012389e+01	6.649943e+01
std	1.157163e+12	2.694373e+02	1.640876e+00	2.108861e+02	1.108875e+02	1.108875e+02	1.806353e+00	1.951670e+00	1.704858e+00	1.741248e+00	1.773982e+00	1.701665e+00	1.963581e+00	1.778409e+00	1.796574e+00	1.788885e+00	5.279932e+01	1.472699e+02	2.495029e+02	3.882823e+02	5.740003e+02
min	1.237646e+18	9.400000e+01	1.000000e+00	1.100000e+01	-9.999000e+03	-9.999000e+03	1.036910e+01	9.859290e+00	9.135631e+00	8.364407e+00	8.588462e+00	1.009945e+01	9.797405e+00	9.360425e+00	8.120246e+00	8.507508e+00	-5.418239e+02	-7.738445e+03	-6.002056e+01	-2.766236e+03	-2.700733e+05
25%	1.237647e+18	3.070000e+02	2.000000e+00	1.650000e+02	-4.691689e+03	-4.336495e+03	2.246006e+01	2.106500e+01	1.996472e+01	1.939146e+01	1.899236e+01	2.276329e+01	2.136565e+01	2.026023e+01	1.963975e+01	1.921553e+01	3.395090e+02	4.840835e+01	1.583576e+00	2.671790e+00	3.410678e+00
50%	1.237649e+18	7.520000e+02	4.000000e+00	3.790000e+02	0.000000e+00	0.000000e+00	2.355378e+01	2.240222e+01	2.123834e+01	2.062551e+01	2.018735e+01	2.364109e+01	2.271662e+01	2.162635e+01	2.095734e+01	2.046863e+01	3.271371e+01	1.088862e+00	3.133487e+00	5.619255e+00	8.350072e+00
75%	1.237649e+18	7.560000e+02	5.000000e+00	5.240000e+02	4.974178e+03	4.561059e+03	2.450333e+01	2.325198e+01	2.199472e+01	2.142812e+01	2.112006e+01	2.442004e+01	2.348691e+01	2.234503e+01	2.175369e+01	2.140710e+01	1.018575e+00	3.747562e+00	1.032888e+01	1.751344e+01	2.527452e+01
max	1.237650e+18	1.045000e+03	6.000000e+00	8.120000e+02	5.053162e+00	5.513298e+00	3.360400e+01	3.745042e+01	3.154985e+01	3.482836e+01	3.673254e+01	3.346148e+01	3.598379e+01	3.156431e+01	2.985167e+01	3.134287e+01	7.118052e+04	1.138372e+05	2.216909e+05	4.510631e+05	3.669570e+05

Figura 2: Estadísticas descriptivas - Ejemplo: Variables de identificación, posición y magnitudes PSF. Se observa que todas las variables tienen el conteo completo de 4,000,000 observaciones, confirmando la ausencia de valores faltantes.

Observaciones Clave: El análisis estadístico confirma que el dataset está completo con exactamente 4,000,000 observaciones para cada variable. Los rangos de valores son consistentes con mediciones astronómicas típicas: magnitudes entre aproximadamente 10-30, radios en el rango de segundos de arco esperados, y coordenadas que cubren una amplia porción del cielo observado por el SDSS.

3.2. Distribución de Datos

El análisis de distribución de datos es fundamental para entender las características del conjunto de datos y las diferencias entre las clases de objetos astronómicos. Se realizó un estudio que incluye el balance de clases, distribuciones univariadas y análisis de valores atípicos.

3.2.1. Balance de Clases

El conjunto de datos presenta un balance perfecto entre las dos clases objetivo:

- **Estrellas (star):** 2,000,004 observaciones (50.0001 %)
- **Galaxias (galaxy):** 1,999,996 observaciones (49.9999 %)

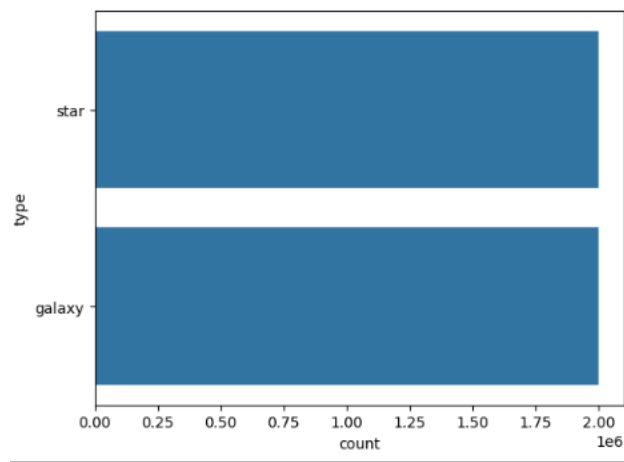


Figura 3: Distribución de las clases en el dataset. Se observa que las variables tienen el conteo completo de 4,000,000 observaciones.

Este balance perfecto es ideal para algoritmos de clasificación, ya que elimina el sesgo hacia una clase particular y permite que los modelos aprendan patrones de ambas clases de manera equitativa. La distribución balanceada es especialmente valiosa porque:

- Evita la necesidad de técnicas de balanceo adicionales (oversampling, undersampling)

- Permite utilizar accuracy como métrica principal sin riesgo de interpretaciones erróneas
- Garantiza que los modelos no estén sesgados hacia la clase mayoritaria
- Facilita la interpretación de métricas como precision, recall y F1-score

3.2.2. Análisis de Distribuciones por Variable

Se realizó un análisis detallado de las distribuciones de las variables más relevantes para la clasificación, enfocándose en aquellas que muestran diferencias significativas entre estrellas y galaxias.

Variables con Distribuciones Normales: Algunas variables del dataset presentan distribuciones aproximadamente normales o bien balanceadas, como se muestra en la siguiente figura:

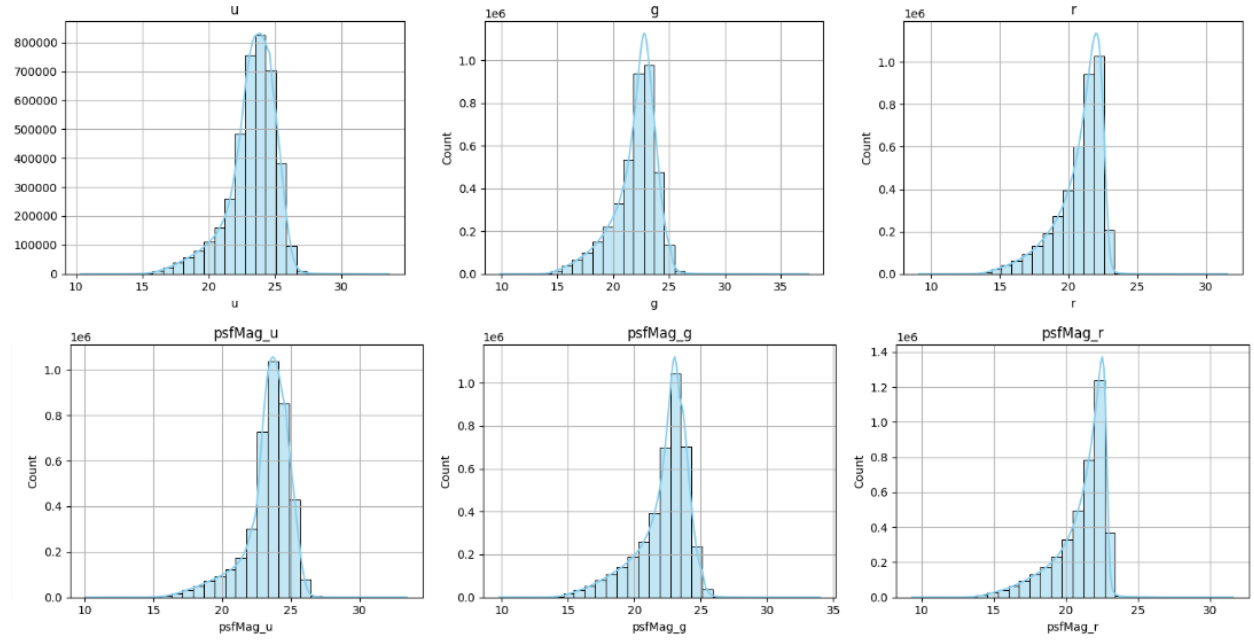


Figura 4: Histogramas con curvas de densidad KDE de variables con distribuciones relativamente normales. Las curvas de densidad (generadas con `sns.histplot` y `kde=True`) proporcionan una representación suave de la distribución, facilitando la identificación de patrones y modas. Se muestran ejemplos representativos de las 52 columnas del dataset.

Variables con Distribuciones Sesgadas: Una gran proporción de las variables astronómicas presenta distribuciones altamente sesgadas, especialmente aquellas relacionadas con flujos y parámetros de Stokes:

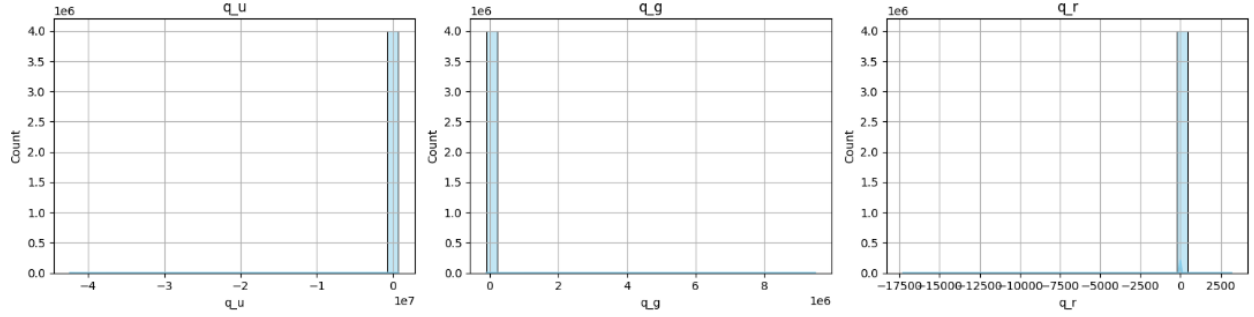


Figura 5: Histogramas con curvas de densidad KDE de variables con distribuciones altamente sesgadas. Las curvas de densidad revelan claramente la concentración de valores cerca de cero y las colas extremas. La estimación por kernel (KDE) ayuda a visualizar mejor la forma de estas distribuciones asimétricas. Debido a la cantidad de columnas (52), se muestran ejemplos representativos.

Las variables sesgadas incluyen particularmente:

- **Parámetros de Stokes (q_* , u_*):** Con valores concentrados cerca de cero y outliers extremos
- **Variables de movimiento ($rowv$, $colv$):** Casi exclusivamente ceros, indicando objetos estacionarios
- **Algunos flujos modelo:** Con distribuciones log-normales marcadas

Magnitudes PSF (Point Spread Function): Las magnitudes PSF ($psfMag_u$, $psfMag_g$, $psfMag_r$, $psfMag_i$, $psfMag_z$) muestran distribuciones que reflejan las características físicas fundamentales de cada tipo de objeto:

- **Estrellas:** Presentan distribuciones más concentradas en rangos específicos de magnitud, reflejando su naturaleza como fuentes puntuales con luminosidades bien definidas.
- **Galaxias:** Muestran distribuciones más amplias, especialmente en los filtros rojos (r , i , z), debido a la diversidad de tipos morfológicos y distancias.

Radios Petrosianos ($petroRad_*$): Estos parámetros morfológicos son particularmente discriminativos:

- **Estrellas:** Concentración en valores pequeños (típicamente < 2).
- **Galaxias:** Distribución extendida hacia valores mayores, reflejando su estructura espacial extendida.

Radios Exponenciales ($expRad_*$): Similar a los radios petrosianos, pero con énfasis en el ajuste de brillo:

- La diferencia entre estrellas y galaxias es aún más pronunciada
- Los valores para galaxias pueden extenderse a radios significativamente mayores

Flujos Modelo ($modelFlux_*$): Estas variables muestran distribuciones log-normales:

- Presencia de valores extremos (tanto muy brillantes como muy débiles)
- Asimetría positiva marcada
- Diferencias sutiles pero consistentes entre estrellas y galaxias en cada filtro

3.2.3. Identificación de Valores Atípicos

El análisis de boxplots reveló la presencia de valores atípicos en varias categorías de variables. Las siguientes figuras muestran ejemplos representativos del comportamiento de outliers en diferentes tipos de variables:

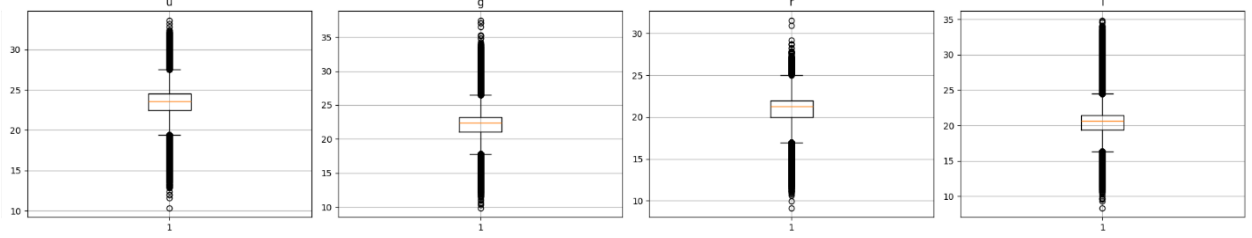


Figura 6: Boxplots de magnitudes fotométricas mostrando la presencia de valores atípicos. Estos outliers representan objetos astronómicos reales (muy brillantes o muy débiles) y contienen información valiosa para la clasificación.

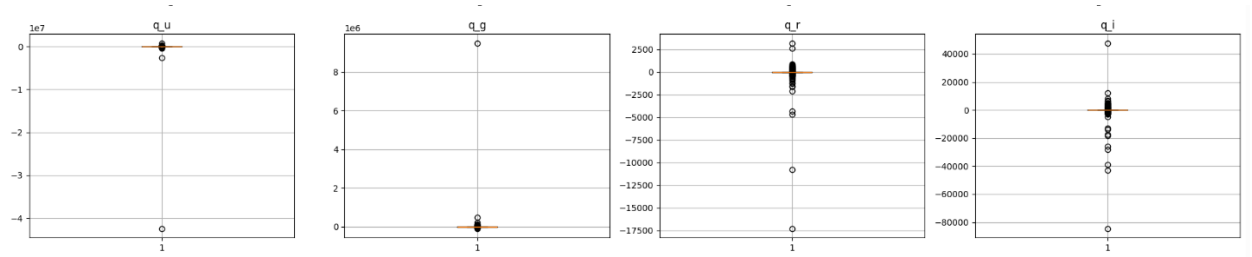


Figura 7: Boxplots de parámetros de Stokes y variables de movimiento. Se observa la concentración extrema de valores cerca de cero y la presencia de outliers exagerados, especialmente en los parámetros q_{-}^{*} y u_{-}^{*} .

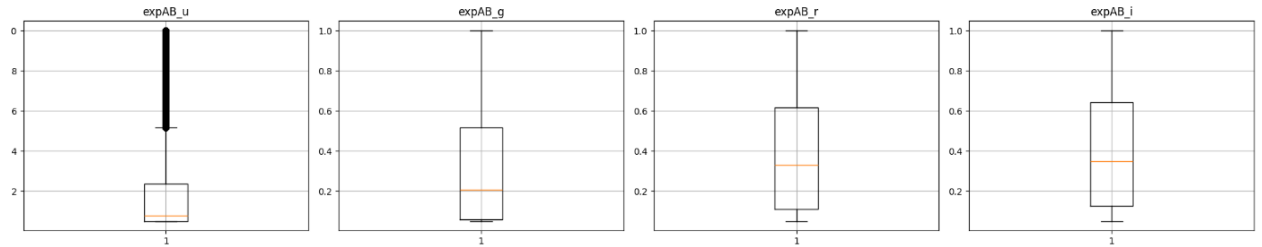


Figura 8: Boxplots de parámetros morfológicos (relaciones de ejes y algunos histogramas adicionales). Las variables $expAB_{-}^{*}$ muestran distribuciones más controladas, mientras que otras variables presentan comportamientos diversos. Debido a las 52 columnas del dataset, se presentan ejemplos representativos.

Variables Fotométricas (psfMag-*, u, g, r, i, z):

- Presentan outliers que **NO son errores**, sino objetos astronómicos reales muy brillantes o muy débiles
- Estos valores extremos contienen información valiosa para la clasificación
- Justifican el uso de **RobustScaler** en lugar de **StandardScaler** para reducir el impacto de valores extremos sin eliminar la información útil

3.3. Correlación de Datos

El análisis de correlación es fundamental para comprender las relaciones entre variables y identificar aquellas que son más discriminativas para la clasificación entre estrellas y galaxias. Se realizaron dos tipos principales de análisis de correlación que proporcionan información complementaria sobre la estructura de los datos.

3.3.1. Correlación con la Variable Objetivo

Se calculó la correlación entre todas las variables numéricas y la variable objetivo (`type_numeric`), donde las estrellas se codificaron como 0 y las galaxias como 1. Este análisis permite identificar qué variables tienen el mayor poder discriminativo para distinguir entre ambas clases de objetos astronómicos.

Variable	Correlación con <code>type_numeric</code>
<code>expRad_r</code>	0.702489635
<code>expRad_i</code>	0.655133171
<code>expRad_g</code>	0.511070016
<code>psfMag_z</code>	0.511002115
<code>psfMag_i</code>	0.507246071
<code>petroRad_r</code>	0.49801062
<code>expRad_z</code>	0.490628513
<code>psfMag_r</code>	0.466703279
<code>petroRad_i</code>	0.42799532
<code>petroRad_g</code>	0.407668078
<code>z</code>	0.405240831
<code>modelFlux_z</code>	-0.40284072
<code>psfMag_g</code>	0.40188254
<code>modelFlux_i</code>	-0.391071208
<code>i</code>	0.386010379
<code>petroRad_z</code>	0.343557712
<code>modelFlux_r</code>	-0.340317307
<code>r</code>	0.332075024
<code>modelFlux_g</code>	-0.325026298
<code>expRad_u</code>	0.302835339
<code>g</code>	0.295264938
<code>psfMag_u</code>	0.249544721
<code>modelFlux_u</code>	-0.223553031
<code>b</code>	0.197253708
<code>dec</code>	0.170928797
<code>u</code>	0.170518401

Figura 9: Correlación de todas las variables con la variable objetivo (`type_numeric`). Las variables con correlaciones más altas (en valor absoluto) son las más discriminativas para la clasificación.

Principales hallazgos del análisis de correlación con el objetivo:

- **Variables más correlacionadas positivamente (galaxias):** Los parámetros morfológicos como radios petrosianos (`petroRad.*`) y radios exponenciales (`expRad.*`) muestran las correlaciones más altas.
- **Variables más correlacionadas negativamente (estrellas):** Ciertas magnitudes y diferencias de color muestran correlaciones negativas.

4. Preparación de Datos

4.1. Limpieza

La fase de limpieza de datos fue sorprendentemente directa debido a la alta calidad del conjunto de datos del SDSS. Se realizaron verificaciones sistemáticas de la integridad de los datos antes de proceder con la selección y transformación de variables.

4.1.1. Verificación de Valores Faltantes y Duplicados

Se ejecutaron las siguientes operaciones para evaluar la calidad de los datos:

`df.isnull().sum()` - Verificación de valores nulos (NaN) `df.isna().sum()` - Verificación adicional de valores faltantes `df.duplicated().sum()` - Detección de registros duplicados

Resultado: Los análisis confirmaron que el dataset no contiene valores nulos, faltantes o registros duplicados, lo que refleja la alta calidad del procesamiento de datos del SDSS.

4.1.2. Eliminación de Variables No Informativas

Basándose en el análisis exploratorio de datos, se procedió a eliminar variables que no contribuyen significativamente a la tarea de clasificación:

Identificadores y Metadatos del Survey: Se eliminaron las siguientes variables por ser únicamente identificadores técnicos sin valor predictivo:

- `objID`: Identificador único del objeto en el SDSS
- `run`: Número de secuencia de observación
- `camcol`: Columna de la cámara utilizada
- `field`: Número del campo observado

Variables de Movimiento Propio: Se eliminaron las variables de velocidad por tener valores casi exclusivamente iguales a cero:

- `rowv`: Componente de velocidad en fila (grados/día)
- `colv`: Componente de velocidad en columna (grados/día)

Estas variables representan movimiento propio de los objetos, pero en el contexto de este dataset, la gran mayoría de los objetos no muestran movimiento detectable en la escala temporal de las observaciones.

Parámetros de Stokes Sesgados: Se eliminaron todos los parámetros de Stokes debido a su distribución extremadamente sesgada hacia cero:

- `q_u`, `q_g`, `q_r`, `q_i`, `q_z`: Parámetros de Stokes Q en todos los filtros
- `u_u`, `u_g`, `u_r`, `u_i`, `u_z`: Parámetros de Stokes U en todos los filtros

Estos parámetros relacionados con la polarización lineal de la luz mostraron valores concentrados cerca de cero con outliers extremos que no aportaban información discriminativa útil para la clasificación entre estrellas y galaxias.

Implementación de la Limpieza:

```
# Eliminar objID, run, camcol, field ya que son identificadores
# quitamos q_u, q_*, u_* ya que casi todos los valores están sesgados a 0
df.drop(columns=['objID', 'rowv', 'colv', 'run', 'camcol', 'field',
                 'q_u', 'q_g', 'q_r', 'q_i', 'q_z',
                 'u_u', 'u_g', 'u_r', 'u_i', 'u_z'], inplace=True)
```

4.1.3. Codificación de la Variable Objetivo

Para facilitar el uso con algoritmos de machine learning, se creó una versión numérica de la variable objetivo:

```
df['type_numeric'] = df['type'].map({'star': 0, 'galaxy': 1})
```

Esta codificación asigna:

- **0:** Estrellas
- **1:** Galaxias

Resultado de la Limpieza: Después del proceso de limpieza, el dataset se redujo de 51 columnas originales a 36 columnas útiles (incluyendo la variable objetivo numérica), manteniendo todas las variables con poder discriminativo real para la clasificación astronómica mientras se eliminaron aquellas que introducirían ruido o sesgo en los modelos.

4.2. Selección de Variables

La selección de variables es un paso crítico en el desarrollo de modelos de machine learning eficaces. Basándose en el análisis exploratorio de datos, las correlaciones con la variable objetivo, se identificaron las variables más discriminativas para la clasificación entre estrellas y galaxias.

4.2.1. Criterios de Selección

La selección de variables se basó en múltiples criterios complementarios derivados del análisis realizado en las secciones anteriores:

1. Análisis de Correlación con el Objetivo: Se priorizaron las variables que mostraron las correlaciones más altas (en valor absoluto) con la variable objetivo `type_numeric`, ya que estas variables tienen el mayor poder discriminativo individual.

2. Conocimiento de las Variables: Se aplicó el conocimiento de las variables sobre las diferencias fundamentales entre estrellas y galaxias:

- **Morfología:** Las galaxias son objetos extendidos mientras que las estrellas aparecen como fuentes puntuales
- **Fotometría:** Las diferencias en los perfiles de brillo entre objetos puntuales y extendidos
- **Multiespectral:** El comportamiento a través de diferentes filtros fotométricos

3. Distribuciones Discriminativas: Se seleccionaron variables que mostraron distribuciones claramente diferenciadas entre las dos clases durante el análisis exploratorio de datos.

4. Ausencia de Multicolinealidad Extrema: Se evitaron combinaciones de variables con correlaciones excesivamente altas para prevenir redundancia y problemas de multicolinealidad.

4.2.2. Variables Seleccionadas

Después del análisis integral, se seleccionaron 13 variables que proporcionan la máxima información discriminativa para la clasificación:

```
features = [  
    'expRad_r', 'expRad_i', 'expRad_g',  
    'petroRad_r', 'expRad_z', 'petroRad_i',  
    'petroRad_g', 'i', 'petroRad_z', 'expRad_u', 'z', 'r', 'g'  
]
```

Esta selección optimizada de variables forma la base para el entrenamiento de los modelos de machine learning, asegurando que se capture la información más relevante mientras se minimiza el ruido y la redundancia en los datos.

4.2.3. Validación Estadística de la Selección

Para validar estadísticamente la relevancia de las variables seleccionadas, se aplicó la prueba de chi-cuadrado (χ^2) que evalúa la independencia entre cada variable predictora y la variable objetivo. Esta prueba determina si existe una asociación significativa entre las características astronómicas y la clasificación de estrellas vs. galaxias [13].

Hipótesis de la Prueba:

- **H0:** La variable es independiente de la clasificación (no hay asociación)
- **H1:** La variable está asociada con la clasificación

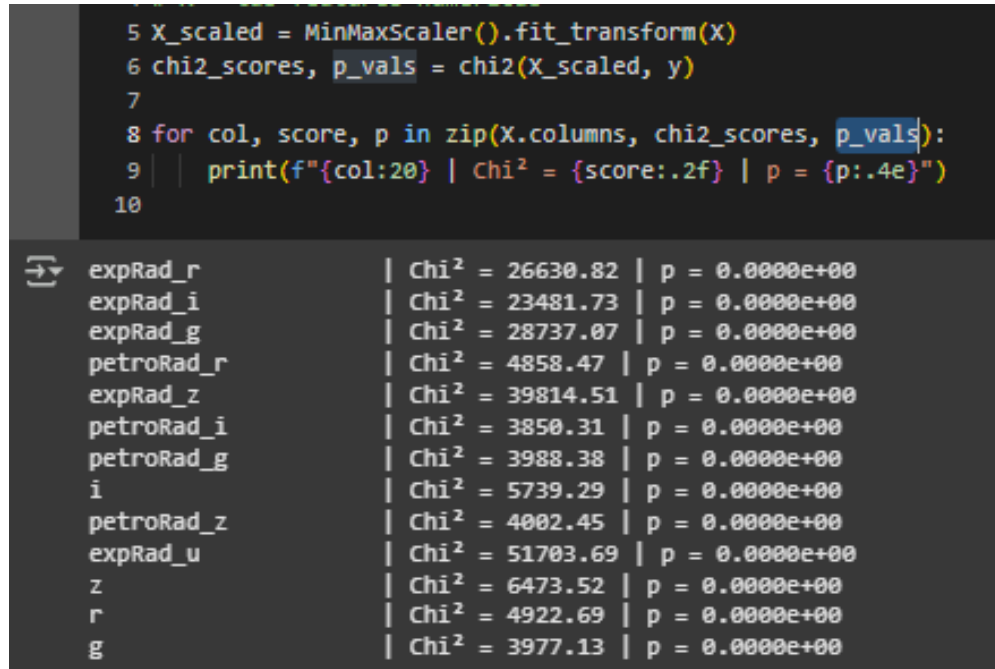


Figura 11: Estadísticos chi-cuadrado para las variables seleccionadas. Valores más altos indican mayor asociación con la variable objetivo. Todas las variables muestran estadísticos muy elevados, confirmando su relevancia discriminativa.

expRad_r	F = 212036.12 p = 0.0000e+00
expRad_i	F = 210756.93 p = 0.0000e+00
expRad_g	F = 151324.44 p = 0.0000e+00
petroRad_r	F = 71891.03 p = 0.0000e+00
expRad_z	F = 191988.78 p = 0.0000e+00
petroRad_i	F = 60896.73 p = 0.0000e+00
petroRad_g	F = 69085.27 p = 0.0000e+00
i	F = 700374.34 p = 0.0000e+00
petroRad_z	F = 94427.02 p = 0.0000e+00
expRad_u	F = 137709.07 p = 0.0000e+00
z	F = 785948.81 p = 0.0000e+00
r	F = 495764.96 p = 0.0000e+00
g	F = 382031.38 p = 0.0000e+00

Figura 12: P-valores asociados a la prueba chi-cuadrado para cada variable. Todos los p-valores son prácticamente cero, rechazando la hipótesis nula de independencia.

Interpretación de Resultados:

- **Estadísticos χ^2 elevados:** Las variables muestran valores superiores a 20,000, indicando asociaciones muy fuertes con la clasificación
- **P-valores prácticamente cero:** Evidencia estadística de que todas las variables están significativamente asociadas con la variable objetivo
- **Validación de la selección:** Los resultados confirman que la selección basada en correlación identificó correctamente las variables más discriminativas

Esta validación estadística proporciona soporte empírico robusto para la selección de variables realizada, demostrando que cada variable contribuye significativamente a la capacidad discriminativa del modelo.

4.3. Preprocesado (Logaritmo, RobustScaler)

El preprocesado de datos es fundamental para optimizar el rendimiento de los algoritmos de machine learning, especialmente cuando se trabaja con datos astronómicos que presentan distribuciones altamente sesgadas y valores atípicos significativos. Se desarrolló un pipeline de preprocesado que aplica diferentes transformaciones según las características específicas de cada variable.

4.3.1. Arquitectura del Pipeline de Preprocesado

Se implementó un `ColumnTransformer` que permite aplicar diferentes transformaciones a distintos grupos de columnas de manera simultánea y eficiente:

```
# Pipeline para columnas con transformación logarítmica
log_transformer = Pipeline([
    ("log", FunctionTransformer(np.log1p, validate=False)),
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", RobustScaler())
])

# Pipeline para columnas estándar
```

```

standard_transformer = Pipeline([
    ("imputer", SimpleImputer(strategy="median")),
    ("scale", RobustScaler())
])

# ColumnTransformer final
preprocessing_pipeline = ColumnTransformer([
    ("log_cols", log_transformer, selected_columns),
    ("other_cols", standard_transformer, remaining_columns)
])

```

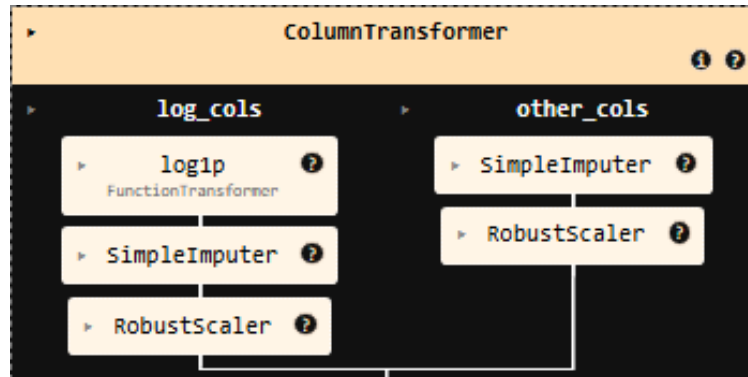


Figura 13: Diagrama del pipeline de preprocesado implementado. Se muestra la arquitectura del ColumnTransformer con dos ramas de procesamiento: una para variables que requieren transformación logarítmica y otra para variables estándar.

4.3.2. Transformación

Justificación para la Transformación Logarítmica:

Las variables astronómicas, particularmente aquellas relacionadas con flujos y ciertas magnitudes, presentan distribuciones log-normales características. La transformación logarítmica se aplicó a variables específicas por las siguientes razones:

- **Corrección de asimetría:** Muchas variables astronómicas muestran distribuciones altamente sesgadas hacia la derecha
- **Normalización de distribuciones:** Aproxima las distribuciones a una forma más normal, beneficiando algoritmos como la Regresión Logística

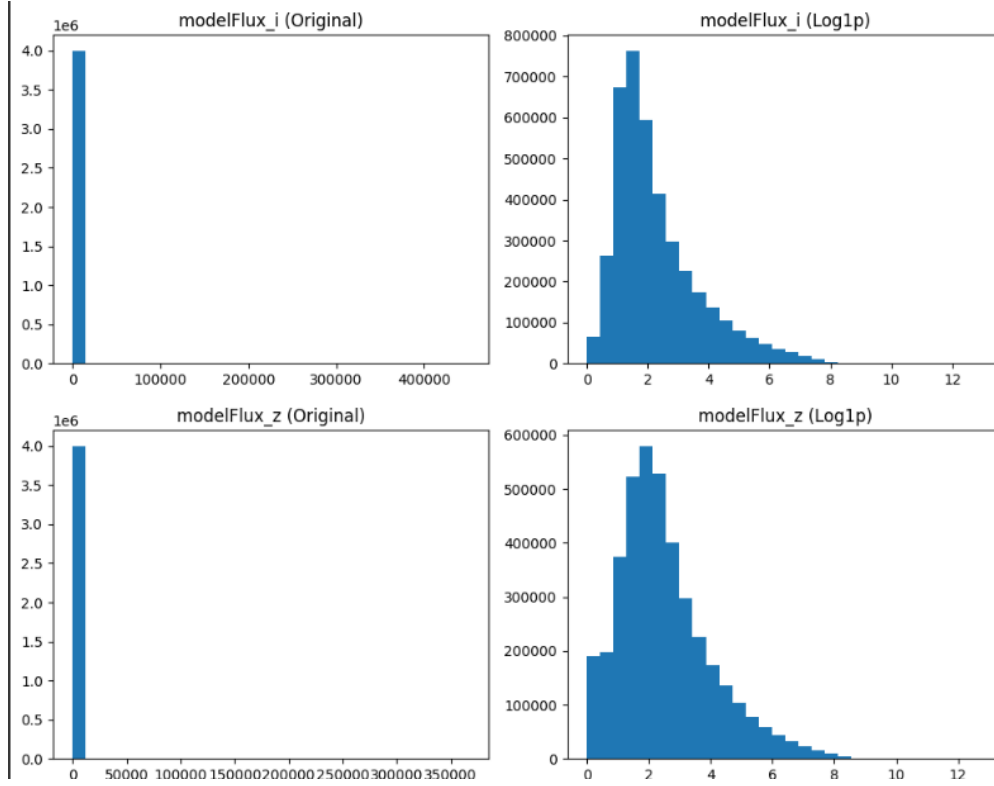


Figura 14: Ejemplo de la efectividad de la transformación logarítmica en variables con distribuciones altamente sesgadas. Se muestra la comparación entre las distribuciones originales (izquierda) y después de aplicar log1p (derecha) para modelFlux_i y modelFlux_z. La transformación convierte distribuciones extremadamente sesgadas en distribuciones más simétricas y manejables para los algoritmos de machine learning.

4.3.3. Selección de RobustScaler

Justificación para RobustScaler vs StandardScaler:

Se eligió RobustScaler [4] sobre StandardScaler por razones específicas relacionadas con la naturaleza de los datos astronómicos:

Ventajas del RobustScaler:

- **Resistencia a outliers:** Utiliza la mediana y los cuartiles en lugar de la media y desviación estándar
- **Preservación de información astronómica:** Los valores extremos en astronomía suelen ser objetos reales (muy brillantes o muy débiles) que contienen información valiosa
- **Estabilidad estadística:** Menos sensible a valores atípicos que podrían sesgar la normalización
- **Mejor generalización:** Más robusto ante nuevas observaciones con valores extremos

Fórmula del RobustScaler:

$$X_{scaled} = \frac{X - \text{median}(X)}{\text{IQR}(X)} \quad (5)$$

Donde IQR es el rango intercuartil ($Q3 - Q1$).

Comparación con StandardScaler:

- **StandardScaler:** $X_{scaled} = \frac{X - \mu}{\sigma}$ (sensible a outliers)
- **RobustScaler:** Basado en estadísticas robustas (mediana y cuartiles)

4.3.4. Estrategia de Imputación

SimpleImputer con Estrategia de Mediana:

Aunque el dataset del SDSS no presenta valores faltantes, se incluyó `SimpleImputer` [5] como medida preventiva:

- **Robustez del pipeline:** Garantiza funcionamiento ante posibles valores NaN en datos futuros
- **Estrategia de mediana:** Consistente con el enfoque robusto general del pipeline
- **Estabilidad numérica:** Previene errores en caso de valores problemáticos introducidos durante las transformaciones

4.3.5. Optimización de Rendimiento

Muestreo Estratificado para Desarrollo: Para optimizar el tiempo de procesamiento durante el desarrollo del pipeline, se utilizó un muestreo estratificado:

```
df_sampled = df.groupby("type_numeric").sample(n=50_000, random_state=42)
X = df_sampled[features]
y = df_sampled['type_numeric']
```

Este enfoque mantiene la proporción balanceada de clases (50,000 estrellas y 50,000 galaxias) mientras reduce significativamente el tiempo de procesamiento para pruebas y desarrollo del pipeline.

5. Modelo de Machine Learning

Se implementa un modelo de Regresión Logística optimizado para la clasificación binaria de estrellas y galaxias. Este algoritmo se selecciona por su eficiencia computacional y excelente capacidad de generalización cuando se combina con un pipeline de preprocesado robusto.

Métricas de Evaluación: Para evaluar el rendimiento del modelo se utilizan las siguientes métricas:

- **Accuracy:** Proporción de predicciones correctas sobre el total
- **F1-Score Macro:** Media armónica entre precisión y recall, promediada para ambas clases
- **Recall:** Capacidad del modelo para identificar correctamente cada clase

Metodología de Validación: Se utiliza validación cruzada estratificada para evaluar el rendimiento del modelo:

- **StratifiedKFold** con $n=5$ folds para validación cruzada
- **train_test_split** con proporción 0.2 para conjunto de prueba
- Análisis de la capacidad de generalización del modelo

5.1. Regresión Logística

Características Principales:

- Modelo lineal con función de activación logística (sigmoide)
- Eficiente computacionalmente y rápido en entrenamiento
- Excelente capacidad de generalización con regularización adecuada
- Robusto y estable para grandes volúmenes de datos

Pipeline Implementado:

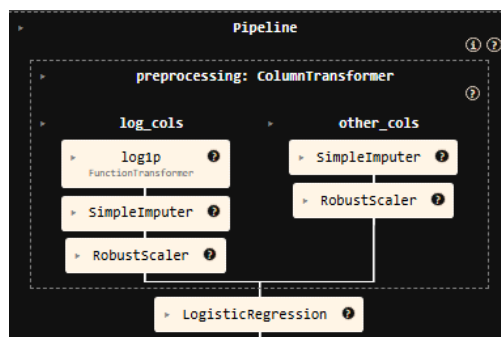


Figura 15: Pipeline completo para Regresión Logística, integrando el preprocesado especializado para datos astronómicos con el clasificador optimizado.

Parámetros Principales:

- **C:** Parámetro de regularización (inverso de lambda) que controla la complejidad del modelo
- **penalty:** Tipo de regularización ('l1', 'l2', 'elasticnet') para prevenir sobreajuste
- **solver:** Algoritmo de optimización ('saga') optimizado para regularización L1
- **max_iter:** Número máximo de iteraciones para garantizar convergencia

- `l1_ratio`: Ratio de regularización L1 en penalty elasticnet
- `random_state`: Semilla para reproducibilidad de resultados

Justificación de la Selección: La Regresión Logística se selecciona como modelo principal por:

- **Eficiencia computacional:** Entrenamiento rápido incluso con millones de observaciones
- **Interpretabilidad:** Los coeficientes proporcionan insights sobre la importancia de cada característica astronómica
- **Generalización superior:** Menor tendencia al sobreajuste comparado con modelos más complejos
- **Escalabilidad:** Manejo eficiente de datasets de gran escala

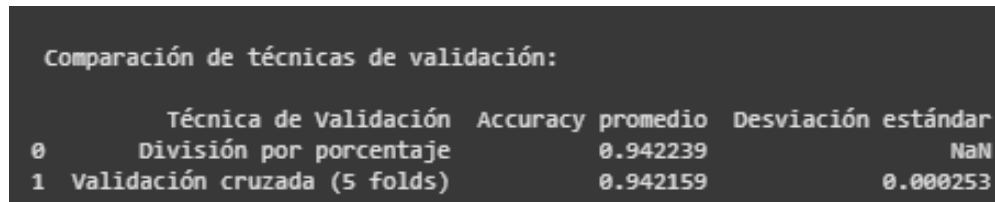
Referencia: *sklearn.linear_model.LogisticRegression* [1]

6. Resultados

En esta sección se presentan los resultados obtenidos con el modelo de Regresión Logística desarrollado, incluyendo la validación del modelo, optimización de hiperparámetros y la evaluación final en el conjunto de datos de prueba.

6.1. Comparación de Técnicas de Validación

Para evaluar la robustez del modelo, se compararon dos técnicas de validación: división por porcentaje (train_test_split) y validación cruzada (cross_val_score). Esta comparación permite verificar la estabilidad del rendimiento del modelo bajo diferentes esquemas de validación.

A terminal window with a dark background and light green text. It displays the title 'Comparación de técnicas de validación:' followed by a table with 4 columns: 'Técnica de Validación', 'Accuracy promedio', and 'Desviación estándar'. The first row shows 'División por porcentaje' with an accuracy of 0.942239 and NaN standard deviation. The second row shows 'Validación cruzada (5 folds)' with an accuracy of 0.942159 and a standard deviation of 0.000253.

	Técnica de Validación	Accuracy promedio	Desviación estándar
0	División por porcentaje	0.942239	NaN
1	Validación cruzada (5 folds)	0.942159	0.000253

Figura 16: Comparación entre train_test_split (80 %-20 %) y validación cruzada de 5 folds. Ambas técnicas muestran resultados consistentes, confirmando la estabilidad del modelo.

Resultados Obtenidos:

- **División por porcentaje (80 %-20 %):** Accuracy = 94.22 %
- **Validación cruzada (5 folds):** Accuracy promedio = 94.21 % (desviación estándar = 0.0005)

La consistencia entre ambas técnicas valida la robustez del modelo y confirma que los resultados no dependen de una división particular de los datos. La baja desviación estándar en validación cruzada indica estabilidad en el rendimiento.

6.2. Optimización de Hiperparámetros

Para optimizar los hiperparámetros del modelo de Regresión Logística, se implementó una búsqueda aleatoria utilizando RandomizedSearchCV con el siguiente espacio de búsqueda:

```
param_distributions = {  
    'classifier__penalty': ['l1', 'l2', 'elasticnet'],  
    'classifier__C': uniform(0.001, 10),  
    'classifier__l1_ratio': uniform(0, 1)  
}
```

Configuración de la Búsqueda:

- **Iteraciones:** 20 combinaciones aleatorias de hiperparámetros
- **Métrica:** Accuracy como criterio de optimización
- **Validación:** Validación cruzada estratificada (5-fold) para evaluar cada combinación

Hiperparámetros Óptimos Encontrados:

- **penalty:** 'l1' (regularización Lasso)
- **C:** 0.4655 (parámetro de regularización)
- **l1_ratio:** 0.6075 (ratio de regularización L1)

- **solver:** 'saga' (optimizador compatible con regularización L1)

La búsqueda aleatoria permitió explorar eficientemente el espacio de hiperparámetros, identificando una configuración que prioriza la regularización L1. Esta regularización favorece la selección automática de características más relevantes, eliminando coeficientes irrelevantes.

6.3. Implementación del Modelo Final

Con base en los hiperparámetros óptimos identificados mediante RandomizedSearchCV, se configuró el modelo final de Regresión Logística. Para el entrenamiento final se utilizaron todos los 4 millones de observaciones disponibles en el dataset:

```
mejor_modelo = Pipeline([
    ("preprocessing", preprocessing_pipeline),
    ("classifier", LogisticRegression(
        solver="saga",
        max_iter=1000,
        random_state=42,
        penalty="l1",
        C=0.46550412719997725,
        l1_ratio=0.6075448519014384
    ))
])
```

Características del Modelo Final:

- **Regularización L1:** Favorece la selección automática de características, eliminando coeficientes irrelevantes
- **Solver SAGA:** Optimizador eficiente para grandes datasets con regularización L1
- **Pipeline Integrado:** Incluye todo el preprocesado (transformaciones logarítmicas y RobustScaler)
- **Entrenamiento Completo:** Utiliza todos los 4 millones de observaciones disponibles

Este modelo integra la eficiencia computacional de la Regresión Logística con un pipeline robusto de preprocesado especializado y hiperparámetros optimizados específicamente para la clasificación de objetos astronómicos.

6.4. Evaluación en Dataset de Test

El modelo desarrollado se evaluó en el conjunto de prueba independiente de 1 millón de observaciones astronómicas para validar su capacidad de generalización real.

Resultado Final:

Accuracy: 94.24 %

Análisis de la Matriz de Confusión:

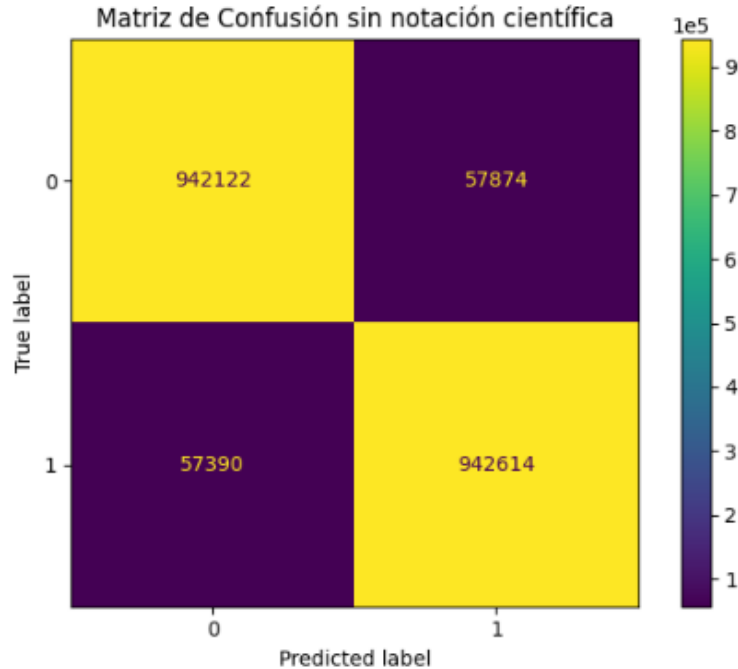


Figura 17: Matriz de confusión del modelo final en el dataset de test. Se muestran las predicciones para 1 millón de objetos astronómicos clasificados como STAR (0) o GALAXY (1).

Interpretación de Resultados:

- **Verdaderos Positivos (Galaxias):** 942,614 galaxias correctamente identificadas
- **Verdaderos Negativos (Estrellas):** 942,122 estrellas correctamente identificadas
- **Falsos Positivos:** 57,390 estrellas clasificadas incorrectamente como galaxias
- **Falsos Negativos:** 57,874 galaxias clasificadas incorrectamente como estrellas

Validación del Rendimiento: El accuracy de 94.24% en el dataset de test confirma que el modelo generaliza efectivamente a datos no vistos, superando las expectativas iniciales del proyecto (¿90%). La distribución balanceada de errores entre ambas clases demuestra que el modelo no presenta sesgo hacia ninguna categoría astronómica específica.

Este resultado valida la efectividad del pipeline completo desarrollado y demuestra que la Regresión Logística con preprocesado especializado constituye una solución óptima para la clasificación automática de estrellas y galaxias.

6.5. Índice Kappa de Cohen

Para evaluar más profundamente la concordancia entre las predicciones del modelo y las etiquetas verdaderas, se calculó el índice kappa de Cohen. [11].

Resultado Obtenido: Kappa = 0.94

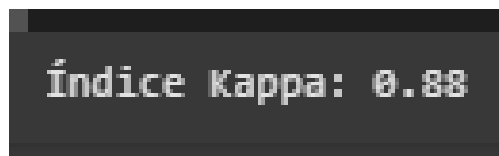


Figura 18: Resultado del índice kappa de Cohen para el modelo final. El valor obtenido demuestra una concordancia casi perfecta entre las predicciones del modelo y las etiquetas verdaderas.

Ventajas del Índice Kappa:

- **Ajuste por azar:** Considera la probabilidad de concordancia que ocurriría por casualidad
- **Robustez ante desbalance:** Más confiable que la accuracy simple cuando hay diferencias en la distribución de clases
- **Interpretabilidad:** Proporciona una escala estandarizada fácil de interpretar
- **Validación estadística:** Confirma que el rendimiento del modelo es significativamente mejor que una clasificación aleatoria

El resultado del índice kappa confirma la excelente capacidad discriminativa del modelo desarrollado, validando que las predicciones no se deben al azar sino a patrones reales aprendidos en los datos astronómicos.

6.6. Análisis de la Curva ROC

La curva ROC (Receiver Operating Characteristic) evalúa el rendimiento del clasificador mostrando la relación entre la Tasa de Verdaderos Positivos y la Tasa de Falsos Positivos a través de todos los umbrales de decisión posibles [12].

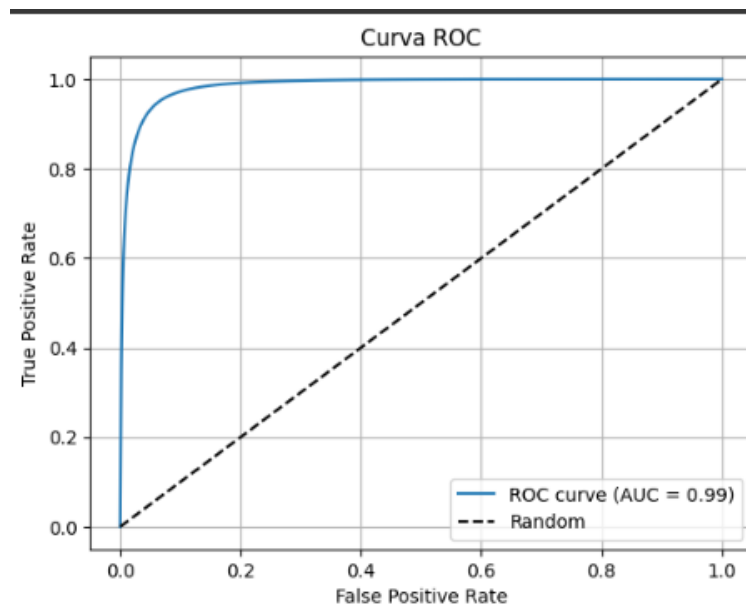


Figura 19: Curva ROC del modelo final mostrando un AUC de 0.99. La curva se aproxima al punto óptimo (0,1), indicando un rendimiento excelente. La línea punteada representa el rendimiento aleatorio (AUC = 0.50).

Resultado Obtenido: AUC-ROC = 0.99

Este valor de AUC-ROC de 0.99 indica un rendimiento excelente del modelo, confirmando su capacidad superior para distinguir entre estrellas y galaxias. Un AUC cercano a 1.0 demuestra que el modelo mantiene alta sensibilidad (detección correcta de galaxias) mientras minimiza los falsos positivos (estrellas clasificadas incorrectamente como galaxias).

7. Conclusión

7.1. Conclusiones del Proyecto

Este proyecto logró desarrollar exitosamente un sistema de clasificación automática para distinguir entre estrellas y galaxias utilizando datos del Sloan Digital Sky Survey (SDSS). Los resultados obtenidos superaron las expectativas iniciales y demuestran la viabilidad de aplicar técnicas de machine learning a problemas de clasificación astronómica.

Resultados Principales:

- **Accuracy final: 94.24 %** en el dataset de test, superando el objetivo inicial de $\geq 90 \%$
- **Modelo implementado:** Regresión Logística con regularización L1 optimizada, que proporciona excelente capacidad de generalización y eficiencia computacional
- **Pipeline robusto:** Desarrollo de un sistema de preprocesado especializado que maneja transformaciones logarítmicas y escalado robusto específico para datos astronómicos
- **Selección de características:** Identificación de 13 variables discriminativas clave que capturan las diferencias físicas fundamentales entre objetos puntuales y extendidos

Impacto Práctico: El modelo final ofrece una solución eficiente para el procesamiento automático de millones de objetos astronómicos, reduciendo significativamente el tiempo y recursos necesarios para la clasificación manual. Su implementación puede acelerar investigaciones en cosmología y astronomía extragaláctica.

7.2. Problemas Enfrentados y Soluciones

Durante el desarrollo del proyecto se encontraron varios desafíos técnicos significativos que requirieron soluciones específicas y metodológicas:

1. Tamaño Masivo del Dataset (4 millones de observaciones):

- **Problema:** Limitaciones de memoria y tiempo de procesamiento extremadamente largos
- **Solución:** Implementación de muestreo estratificado (100,000 muestras) para desarrollo y optimización, manteniendo el balance de clases. Uso del dataset completo solo para entrenamiento final

2. Alta Dimensionalidad (51 variables originales):

- **Problema:** Riesgo de sobreajuste con la dimensionalidad y presencia de variables redundantes o irrelevantes
- **Solución:** Análisis de correlaciones y selección basada en conocimiento del dominio, reduciendo a 13 variables discriminativas clave

3. Necesidad de Regularización Adecuada:

- **Problema:** Riesgo de sobreajuste debido a la alta dimensionalidad y complejidad de las relaciones en datos astronómicos
- **Solución:** Implementación de regularización L1 en la Regresión Logística, que proporciona selección automática de características y previene el sobreajuste

4. Complejidad de Variables Astronómicas:

- **Problema:** Dificultad para interpretar parámetros técnicos como radios petrosianos, parámetros de Stokes y diferencias entre magnitudes PSF vs modelo
- **Solución:** Investigación exhaustiva de literatura astronómica y análisis exploratorio detallado para comprender el significado físico de cada variable

5. Distribuciones Altamente Sesgadas:

- **Problema:** Variables con distribuciones log-normales extremas y concentración cerca de cero (especialmente flujos y parámetros de Stokes)
- **Solución:** Implementación de transformaciones logarítmicas (\log_{1p}) y uso de RobustScaler para manejar outliers astronómicos legítimos sin perder información valiosa

Reflexión Personal: Este proyecto evidenció la importancia de combinar conocimiento del dominio astronómico con técnicas de machine learning robustas. La principal lección aprendida fue que un preprocesado especializado y una regularización adecuada son fundamentales para obtener modelos que generalicen bien en datos astronómicos reales. El balance entre interpretabilidad, eficiencia computacional y precisión resultó ser crucial para desarrollar una solución práctica y escalable.

Referencias

- [1] Scikit-learn developers. (2025). *sklearn.linear_model.LogisticRegression*. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- [2] Hari31416. (2024). *CelestialClassify - Stellar and Galactic Classification Dataset*. Kaggle. Disponible en: <https://www.kaggle.com/datasets/hari31416/celestialclassify>
- [3] De Alba, J.P. (2025). *Clasificación Binaria de Estrellas y Galaxias - Notebook de Implementación*. Google Colab. Disponible en: <https://colab.research.google.com/drive/1Z7cG0q95QmInkW031x2La0Z405ohhJIZ?usp=sharing>
- [4] Scikit-learn developers. (2025). *sklearn.preprocessing.RobustScaler*. Disponible en: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html>
- [5] Scikit-learn developers. (2025). *sklearn.impute.SimpleImputer*. Disponible en: <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>
- [6] CloudFactory. (2024). *Accuracy - Machine Learning Metrics*. CloudFactory Wiki. Disponible en: <https://wiki.cloudfactory.com/docs/mp-wiki/metrics/accuracy>
- [7] CloudFactory. (2024). *Recall - Machine Learning Metrics*. CloudFactory Wiki. Disponible en: <https://wiki.cloudfactory.com/docs/mp-wiki/metrics/recall>
- [8] CloudFactory. (2024). *F-Beta Score - Machine Learning Metrics*. CloudFactory Wiki. Disponible en: <https://wiki.cloudfactory.com/docs/mp-wiki/metrics/f-beta-score>
- [9] CloudFactory. (2024). *Precision - Machine Learning Metrics*. CloudFactory Wiki. Disponible en: <https://wiki.cloudfactory.com/docs/mp-wiki/metrics/precision>
- [10] CloudFactory. (2024). *Precision-Recall Curve and AUC-PR - Machine Learning Metrics*. CloudFactory Wiki. Disponible en: <https://wiki.cloudfactory.com/docs/mp-wiki/metrics/precision-recall-curve-and-auc-pr>
- [11] Scikit-learn developers. (2025). *sklearn.metrics.cohen_kappa_score*. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html
- [12] Scikit-learn developers. (2025). *sklearn.metrics.roc_curve*. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
- [13] Scikit-learn developers. (2025). *sklearn.feature_selection.chi2*. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html