CLASIFICACIÓN DE OBJETOS ASTRONÓMICOS

Estadística Avanzada

Álvaro González González
Diciembre de 2023

Índice

1. Motivación

1.1. Contexto Profesional

El fascinante campo de la astronomía ha experimentado una transformación radical con la llegada de la era de los grandes datos. La cantidad abrumadora de información recogida por telescopios y misiones espaciales ha abierto nuevas fronteras en nuestra comprensión del universo. Sin embargo, este vasto océano de datos plantea un desafío significativo: ¿Cómo podemos clasificar y analizar eficientemente esta información para descubrir nuevos conocimientos? Esta pregunta central motiva mi proyecto.

1.2. Identificación del Problema

En el estudio del cosmos, la clasificación de objetos como estrellas, galaxias y quásares es un reto fundamental. Estos cuerpos, aunque distintos en naturaleza y características, pueden aparecer similares a través de observaciones telescópicas, especialmente cuando se encuentran a grandes distancias.

El desafío principal radica en diferenciar de manera precisa entre estos objetos utilizando datos fotométricos y espectroscópicos[1]. La identificación correcta es esencial para estudios posteriores sobre su composición, distribución y evolución, y para entender mejor el universo en su conjunto. Sin embargo, la similitud en sus señales observacionales y la inmensa cantidad de datos recolectados por telescopios modernos hacen que la clasificación manual sea impracticable.

El objetivo principal de este proyecto es aplicar técnicas avanzadas de aprendizaje automático y análisis estadístico para clasificar estos objetos de manera eficiente y precisa.

2. Recopilación de Datos

Los datos para este estudio se han obtenido del Sloan Digital Sky Survey (SDSS), específicamente de su decimoséptima entrega de datos (DR17)[2]. Este recurso es una extensa base de datos astronómica que ofrece medidas fotométricas y espectroscópicas de una amplia variedad de objetos celestes.

La base de datos del SDSS DR17 es ideal para nuestro estudio debido a su rica colección de información sobre estrellas, galaxias y quásares. Incluye datos como magnitudes en diferentes bandas espectrales, desplazamientos al rojo (redshifts), y clasificaciones espectroscópicas, entre otros. Este conjunto de datos es fundamental para el desarrollo de modelos de clasificación basados en aprendizaje automático, permitiéndonos analizar patrones y características de estos objetos astronómicos con un nivel de detalle sin precedentes.

La recopilación de datos se realizó a través de una consulta SQL específica en la interfaz de búsqueda del SDSS DR17, ejecutada desde la herramienta SQL Search de SkyServer[3]. La consulta fue diseñada para seleccionar atributos relevantes para la clasificación de 100.000 objetos celestes, resultando en un conjunto de datos exportado en formato CSV.

La elección de SDSS DR17 se debe a su reconocida precisión, amplitud y la diversidad de sus datos, lo que lo convierte en una de las bases de datos astronómicas más utilizadas en investigaciones de vanguardia.

3. Análisis Exploratorio de Datos

3.1. Descripción de Variables

El conjunto de datos de clasificación estelar recopilado contiene diversas variables fundamentales para el análisis astronómico y la clasificación de objetos celestes. Las variables incluyen:

- **obj_ID**: Un identificador único para cada objeto observado, importante para el seguimiento de datos pero no utilizado en análisis predictivos.
- **Coordenadas Celestes**: `alpha` (ascensión recta) y `delta` (declinación), que proporcionan la ubicación del objeto en el cielo.
- **Magnitudes Fotométricas**: Las bandas espectrales `u`, `g`, `r`, `i`, `z`, que miden la intensidad de la luz en diferentes longitudes de onda, desde el ultravioleta hasta el infrarrojo lejano.
- **Medidas Espectroscópicas y Observacionales**: Incluyendo `run_ID`, `rerun_ID`, `cam_col`, `field_ID`, `spec_obj_ID`, `plate`, `MJD`, `fiber_ID`. Aunque son importantes para la observación, no son relevantes para el análisis predictivo y por tanto se descartan para la clasificación.

- **class**: La clasificación del objeto, que puede ser galaxia, estrella o quásar (QSO), es nuestra variable objetivo para la clasificación.
- **redshift**: Una medida del desplazamiento al rojo, que indica la velocidad a la que un objeto se aleja y puede ser un indicador de la distancia a la que se encuentra.

Para la clasificación, las magnitudes en las distintas bandas espectrales y el desplazamiento al rojo (`redshift`) son las variables clave debido a su relevancia física directa. Las otras variables, como identificadores y detalles observacionales, aunque importantes para el proceso de recopilación de datos, no contribuyen al modelo predictivo y por lo tanto pueden ser excluidas del análisis. Las coordenadas celestes, `alpha` y `delta`, se mantendrán inicialmente en el análisis para determinar si existe alguna correlación espacial, aunque, bajo la suposición de isotropía del universo a gran escala, no se espera que tengan una influencia significativa en la clasificación de los objetos celestes.

Se ha verificado la ausencia de valores faltantes y duplicados, y se han tratado valores extremos en las magnitudes `u` y `z`. Procederemos a realizar visualizaciones y pruebas estadísticas para explorar las características y relaciones en los datos, estableciendo las bases para el modelado predictivo posterior.

3.2. Estadísticas Descriptivas

En la fase de análisis exploratorio, se calcularon las medidas de tendencia central y dispersión para cada variable numérica. Los resultados se presentan en el Cuadro 1 del apéndice. Este resumen proporciona una visión general de la distribución de cada variable clave en nuestro conjunto de datos. Del Cuadro 1, podemos inferir que:

- La media de las magnitudes en las bandas espectrales varía moderadamente, indicando una amplia gama de brillo entre los objetos estudiados.
- El *redshift* tiene una media de 0.57, pero con un máximo de 7.01, sugiriendo que algunos objetos están significativamente más rojos (y por lo tanto, posiblemente más lejanos o con mayor velocidad de alejamiento) que la mayoría.
- Las desviaciones estándar para las coordenadas celestes *alpha* y *delta* son grandes, reflejando una amplia dispersión espacial en el cielo.
- Los valores mínimos y máximos de las magnitudes muestran una variabilidad significativa, lo que puede ser indicativo de diferentes tipos de objetos astronómicos en el conjunto de datos.

3.3. Visualización de Datos

Se elaboraron gráficos de densidad para las variables fotométricas u, g, r, i, z, el *redshift*, y las coordenadas astronómicas *alpha* y *delta*. Estos gráficos se segmentaron según la clase de objeto: Galaxia, QSO y Estrella, proporcionando una comprensión visual clara de la distribución de estas medidas entre los diferentes tipos de objetos astronómicos. Véanse las Figuras 1, 2, 3, 4, 5, 6, 7 y 8.

Se observó que la distribución del *redshift* estaba dominada por los valores extremadamente bajos. Esto hace que sea difícil discernir cualquier detalle de la distribución, ya que todos los datos se aglomeran en un pico estrecho y alto cerca de cero. Véase la Figura 9

Se aplicó una transformación logarítmica al *redshift* para mejorar la interpretabilidad de los datos.

De los gráficos de densidad podemos inferir lo siguiente:

- Las gráficas de densidad desvelan patrones distintivos por clase de objeto astronómico. Los cuásares, representados en naranja, exhiben picos definidos a través de todas las bandas espectrales, reflejando una emisión consistente y poderosa, típica de su naturaleza altamente energética. Las estrellas, en verde, presentan distribuciones amplias y magnitudes más bajas, lo que indica una variabilidad significativa en tamaño y temperatura. Las galaxias, mostradas en azul, se distinguen por tener dos picos definidos que se intensifican en las bandas hacia el ultravioleta, sugiriendo la presencia de distintos tipos de galaxias, algunas de las cuales pueden estar experimentando activa formación estelar.
- El solapamiento significativo entre las distribuciones de las diferentes clases puede complicar la clasificación basada solo en estas variables. Sin embargo, la posición y la altura de los picos en las distribuciones de cada clase podrían ser relevantes.

- El solapamiento ligero entre las distribuciones de *redshift* de cuásares y galaxias puede indicar regiones donde estas clases de objetos tienen propiedades similares o están a distancias comparables, siendo más lejanos los cuásares en general. Sin embargo, la marcada separación de las estrellas, con valores de *redshift* cercanos a cero, confirma que están significativamente más cercanas, lo cual es coherente con el entendimiento de que las estrellas observadas son principalmente de nuestra propia galaxia y no se están alejando de nosotros a velocidades significativas como las galaxias externas y los cuásares, que observamos a través del efecto Doppler cósmico.
- Las gráficas de densidad de las coordenadas celestes alpha y delta muestran patrones cíclicos para las tres clases, pero no hay diferencias marcadas que sugieran que estas coordenadas sean determinantes para la clasificación.

3.4. Análisis de Correlación de Variables

La matriz de correlación (Figura 10) proporciona los valores del coeficiente de correlación de Pearson entre cada par de variables. El coeficiente de correlación de Pearson es una medida estadística que calcula la relación lineal entre dos variables. Se calcula utilizando la siguiente fórmula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

donde:

- r_{xy} es el coeficiente de correlación entre las variables X e Y.
- x_i y y_i son los valores individuales de las variables X e Y.
- $\bar{x}$ y $\bar{y}$ son las medias de las variables X e Y.

Este coeficiente puede tomar valores desde -1 hasta 1. Un valor cercano a 1 indica una correlación positiva fuerte, lo que significa que cuando una variable aumenta, la otra también tiende a aumentar. Un valor cercano a -1 indica una correlación negativa fuerte, significando que cuando una variable aumenta, la otra tiende a disminuir. Un valor cercano a 0 indica que no hay una relación lineal entre las variables.

Aquí hay algunos puntos clave que se pueden extraer de la matriz de correlación:

- Las bandas fotométricas *(u, g, r, i, z)* tienen correlaciones débiles a moderadas con la variable *class*, lo que indica que hay una relación significativa entre algunas mediciones fotométricas y la clasificación de un objeto astronómico.
- La variable *redshift* muestra la mayor correlación positiva con *class*, lo que implica que el desplazamiento al rojo puede ser un buen predictor del tipo de objeto astronómico (estrella, galaxia, quásar).
- Las coordenadas celestes (ascensión recta y declinación) tienen correlaciones muy bajas con la clase, lo que indica que la posición de un objeto en el cielo no está fuertemente relacionada con sus propiedades fotométricas o su clasificación. Por lo tanto las vamos a descartar de nuestro estudio posterior.

4. Modelado Estadístico

4.1. Regresión Logística Multinomial

4.1.1. Conceptos teóricos

En la regresión logística, se modela la probabilidad de que la variable dependiente categórica pertenezca a una clase específica. Para el caso binario, la probabilidad de que Y sea igual a 1 (dada una observación X) se modela como:

$$P(Y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)$$

donde:

- $\sigma(z) = \frac{1}{1+e^{-z}}$ es la función logística o sigmoide.
- $\beta_0, \beta_1, ..., \beta_n$ son los coeficientes del modelo.
- $X_1, X_2, ..., X_n$ son las variables independientes (características como u, g, r, i, z, redshift).

La función de coste en la regresión logística, conocida como log-loss, se utiliza para cuantificar el error entre las predicciones del modelo y los datos observados. Para la clasificación binaria, la función de coste es:

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\beta(x^{(i)})) + (1-y^{(i)})\log(1-h_\beta(x^{(i)}))]$$

donde:

- m es el número de observaciones.
- $y^{(i)}$ es la clase real de la i-ésima observación.
- $h_\beta(x^{(i)})$ es la probabilidad predicha de que la i-ésima observación pertenezca a la clase 1.

Los coeficientes del modelo β se estiman maximizando la función de verosimilitud (equivalente a minimizar la función de coste). Esto se hace comúnmente usando métodos iterativos como el descenso de gradiente.

En el caso de clasificación multiclase la regresión logística se extiende a la regresión logística multinomial. Aquí, se modela la probabilidad de cada clase mediante un conjunto de ecuaciones logísticas, una para cada clase. En la práctica, se suelen emplear técnicas como "one-vs-rest"(OvR) o "softmax".

Para prevenir el sobreajuste, especialmente en conjuntos de datos con muchas características, se aplican técnicas de regularización como L1 (Lasso) o L2 (Ridge). Estas técnicas añaden un término de penalización a la función de coste, controlando así la magnitud de los coeficientes del modelo.

4.1.2. Implementación en Python

La función `LogisticRegression()` de Scikit-learn implementa el modelo de regresión logística en Python. A continuación se detalla su funcionamiento en el contexto del análisis de clasificación:

1. **Creación del Modelo:** Al invocar `LogisticRegression()`, se crea una instancia del modelo de regresión logística. Esta función tiene varios parámetros que permiten personalizar el modelo, como `solver`, `penalty`, y `C` que son el algoritmo de optimización, el tipo de regularización y el inverso de la fuerza de regularización, respectivamente.
2. **Optimización**: Scikit-learn ofrece varios algoritmos de optimización a través del parámetro `solver`. Estos incluyen `liblinear`, `newton-cg`, `lbfgs`, `sag` y `saga`. Cada uno de estos `solvers` tiene sus características y aplicaciones recomendadas, y algunos de ellos utilizan variantes del método de descenso del gradiente.
3. **Ajuste del Modelo (Entrenamiento):** El método `fit()` se utiliza para entrenar el modelo con los datos de entrenamiento. Durante este proceso, el modelo utiliza los datos de entrada (`X_train`), y las etiquetas de salida (`y_train`) para aprender los coeficientes que mejor predicen la variable dependiente.
4. **Predicción:** Una vez entrenado, el modelo puede hacer predicciones sobre nuevos datos usando el método `predict()`. Esto devuelve la clase predicha para cada observación. Además, el método `predict_proba()` puede proporcionar las probabilidades estimadas para cada clase, lo cual es especialmente útil en clasificaciones multiclase y para evaluar la confianza del modelo en sus predicciones.
5. **Coeficientes del Modelo:** Los coeficientes aprendidos por el modelo (accesibles a través del atributo `coef_` del objeto modelo) representan la importancia relativa de cada característica para predecir cada clase.
6. **Regularización:** Incluye regularización (L2 por defecto) para prevenir el sobreajuste, especialmente útil en situaciones con muchas características.

4.1.3. Elección del modelo

La elección de la regresión logística multinomial para nuestro proyecto de clasificación estelar se basa en su idoneidad para manejar nuestra variable objetivo, *class*, que es categórica y multinomial. Este método no solo asigna categorías sino que también proporciona probabilidades para cada una, ofreciendo una comprensión más profunda de nuestras predicciones. Su capacidad para tratar tanto variables continuas como categóricas lo hace flexible y adecuado para nuestros datos astronómicos. Además, la regresión logística multinomial es eficaz con grandes muestras y evita el sobreajuste mediante la regularización, asegurando la generalización y relevancia de nuestras conclusiones. En resumen, su combinación de precisión, interpretabilidad y flexibilidad la convierte en la elección óptima para nuestro análisis.

5. Evaluación del Ajuste del Modelo

Vamos a evaluar el ajuste del modelo que hemos creado para la clasificación de objetos astronómicos. La evaluación del ajuste de un modelo en aprendizaje automático implica considerar varios aspectos importantes para garantizar que el modelo no solo se ajusta bien a los datos de entrenamiento, sino que también generaliza adecuadamente a nuevos datos. En este contexto, nos enfocaremos en dos aspectos principales: el riesgo de sobreajuste (overfitting) o ajuste insuficiente (underfitting), y la utilización de técnicas de regularización.

5.1. Sobreajuste y Ajuste Insuficiente

- Sobreajuste (Overfitting): Sucede cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando tanto las tendencias subyacentes como el ruido en los datos. Como resultado, puede tener un rendimiento pobre en datos nuevos o no vistos. Indicadores típicos de sobreajuste incluyen una alta precisión en los datos de entrenamiento pero una precisión significativamente más baja en los datos de prueba.

- Ajuste Insuficiente (Underfitting): Ocurre cuando un modelo es demasiado simple para capturar la complejidad de los datos, resultando en un rendimiento deficiente tanto en los datos de entrenamiento como en los de prueba. Esto puede ser indicativo de que el modelo no ha aprendido las tendencias subyacentes en los datos.

Para evaluar si nuestro modelo sufre de sobreajuste o ajuste insuficiente, podemos analizar su curva de aprendizaje. Las curvas de aprendizaje son una herramienta útil para evaluar cómo el rendimiento del modelo cambia a medida que aumenta el tamaño del conjunto de entrenamiento. Nuestra curva de aprendizaje (Figura 11) sugiere que el modelo logra un balance adecuado entre el sesgo y la varianza, ya que las curvas de entrenamiento y validación convergen y mantienen una diferencia mínima y consistente, reflejando un rendimiento robusto sin indicios de sobreajuste. La estabilización de la precisión con el aumento del tamaño del conjunto de entrenamiento indica que el modelo ha capturado las tendencias esenciales de los datos, y obtener más datos posiblemente no mejoraría significativamente su precisión. Además, la alta precisión obtenida (aproximadamente 96.1 %) sugiere que el modelo tiene la complejidad necesaria para captar las relaciones en los datos, sin caer en un ajuste insuficiente.

5.2. Técnicas de Regularización

- Regularización: Para el desarrollo de mi modelo opté por la regularización L2 junto con el `solver` *lbfgs*. La elección de la regularización L2 se debe a su capacidad para manejar eficazmente la multicolinealidad entre las variables predictoras sin eliminarlas del modelo.[4]

- Ajuste de Hiperparámetros: En el proceso de optimización de hiperparámetros, inicialmente empleé `RandomizedSearchCV()` para explorar el espacio de hiperparámetros de manera más eficiente y menos costosa que una búsqueda exhaustiva. Sin embargo, debido al alto coste computacional y al esfuerzo que esto implicaba, me vi en la necesidad de realizar ajustes manuales iterativos. Este enfoque práctico me permitió afinar los parámetros de forma más controlada y atendiendo a la retroalimentación inmediata del rendimiento del modelo.

  Durante este ajuste, noté que la precisión del modelo mejoraba al incrementar el valor del parámetro de regularización C. Mientras que el valor predeterminado en muchas implementaciones es $C = 1$, encontré que $C = 100$ ofrecía un balance óptimo, proporcionando suficiente flexibilidad al modelo para ajustarse a los datos sin caer en el sobreajuste.

  Además, fue necesario incrementar el número de iteraciones máximas (`max_iter`) debido a las advertencias de no convergencia que recibía. Este cambio permitió que el algoritmo de optimización iterara adecuadamente hasta encontrar un conjunto de parámetros bien ajustados, asegurando así la estabilidad y la confiabilidad del modelo final.

6. Resultados

Se ha creado una matriz de confusión (Cuadros 2 y 3) que representa el rendimiento del modelo de clasificación en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada una de las clases (GALAXY, STAR, QSO).

Anexos

Cuadros

Medida	alpha	delta	u	g	r	i	z	redshift
Media	177.28	24.14	22.08	20.63	19.65	19.08	18.76	0.57
Desviación Estándar	96.50	19.64	2.25	2.03	1.85	1.76	1.76	0.73
Mínimo	0.0055	-18.78	10.99	10.49	9.82	9.47	9.61	-0.009
25 %	127.52	5.15	20.35	18.96	18.14	17.72	17.46	0.05
Mediana	180.91	23.65	22.18	20.21	19.41	19.05	18.90	0.42
75 %	233.89	39.92	23.68	22.13	21.05	20.40	19.92	0.70
Máximo	359.99	83.00	32.78	31.60	29.57	32.14	29.38	7.01

Cuadro 1: Resumen estadístico de las variables numéricas.

	Predicción: GALAXY	**Predicción: STAR**	**Predicción: QSO**
Real: GALAXY	17409	160	292
Real: STAR	0	6448	0
Real: QSO	672	1	5018

Cuadro 2: Matriz de confusión para las clases GALAXY, STAR y QSO.

	Predicción: GALAXY (%)	**Predicción: STAR (%)**	**Predicción: QSO (%)**
Real: GALAXY	97.47	0.9	1.63
Real: STAR	0.0	100.0	0.0
Real: QSO	11.81	0.02	88.17

Cuadro 3: Matriz de confusión en porcentajes para las clases GALAXY, STAR y QSO.

Figuras

Figura 1: Densidad de la variable alpha.

Figura 2: Densidad de la variable delta.

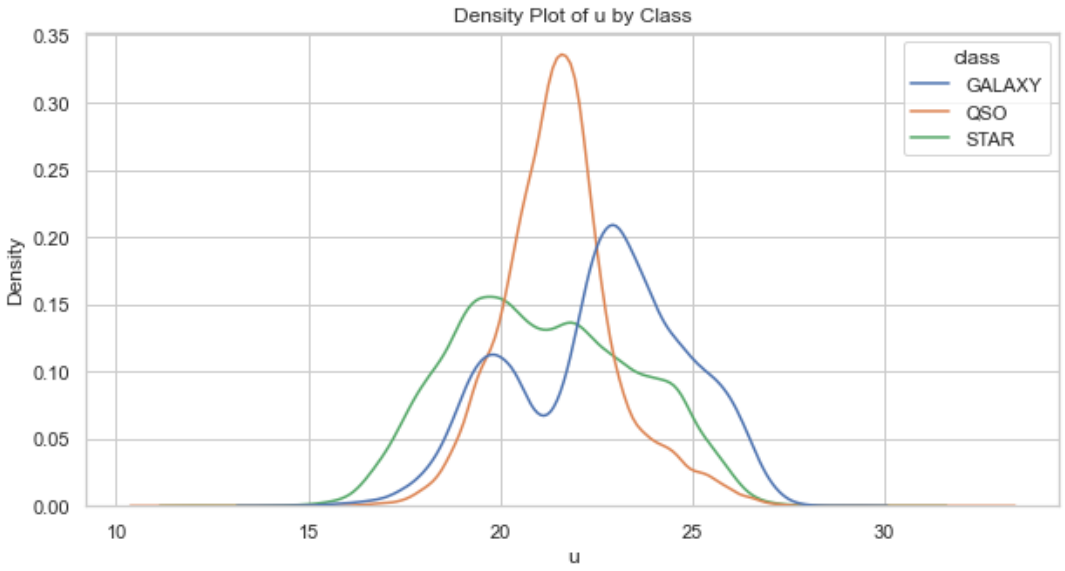

Figura 3: Densidad de la variable u.

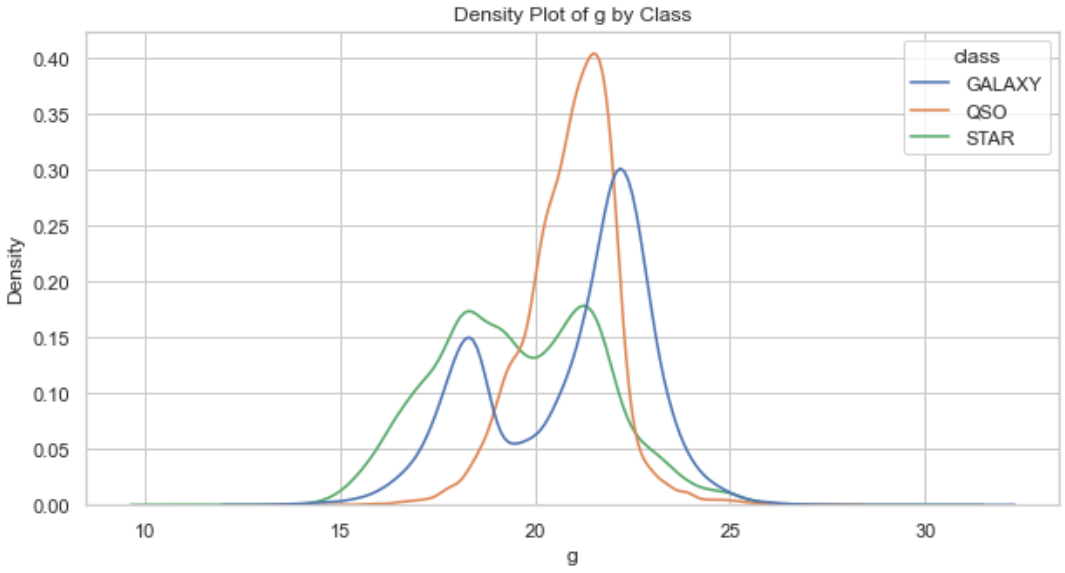

Figura 4: Densidad de la variable g.

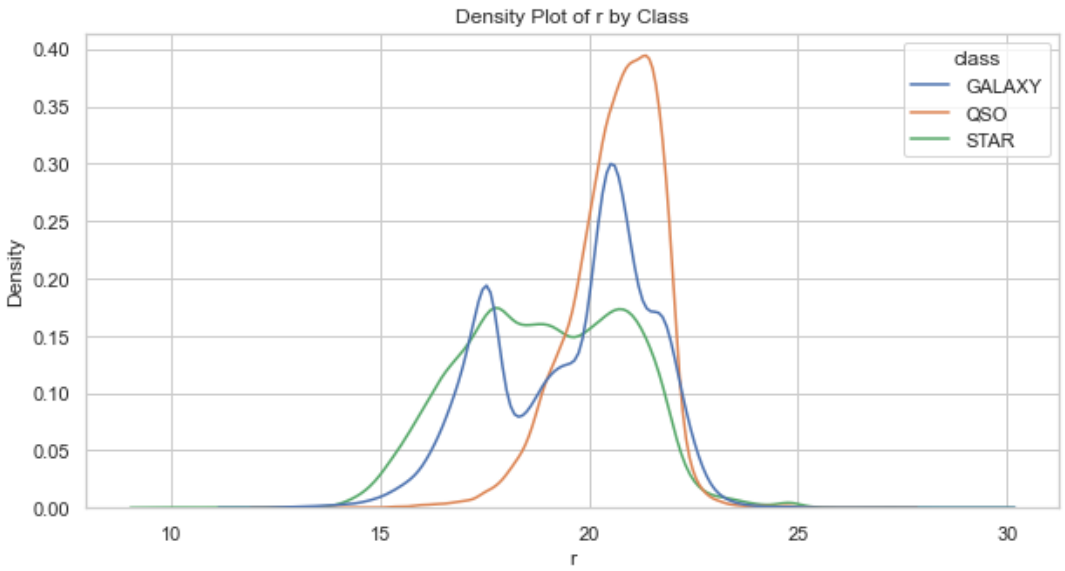

Figura 5: Densidad de la variable r.

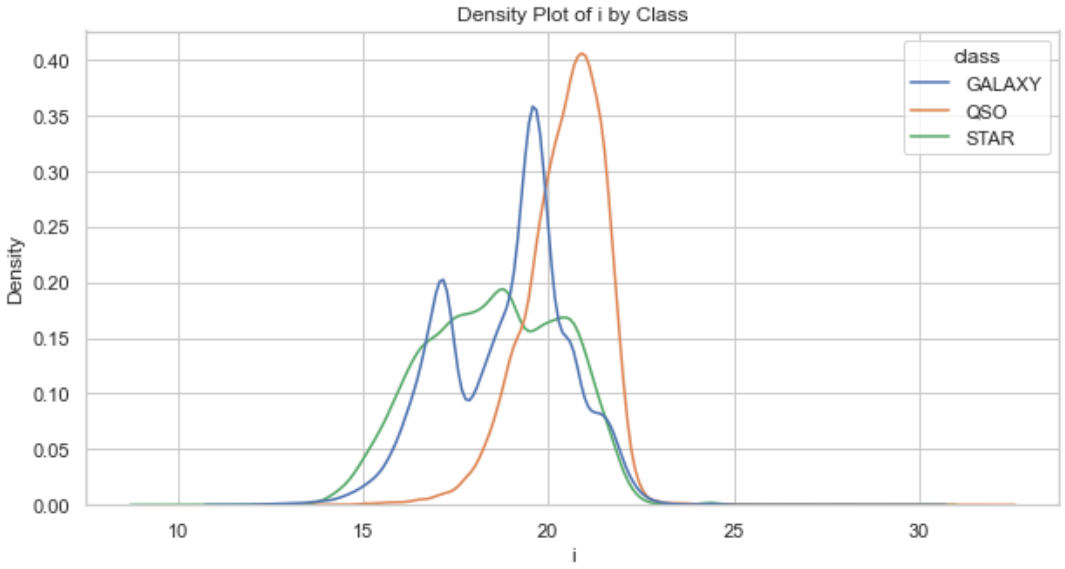

Figura 6: Densidad de la variable i.

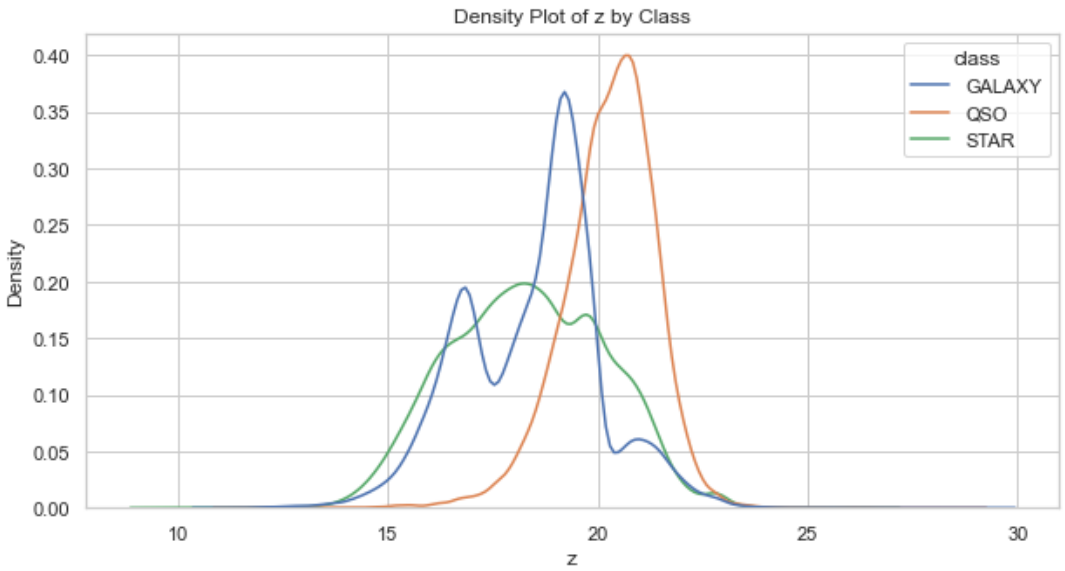

Figura 7: Densidad de la variable z.

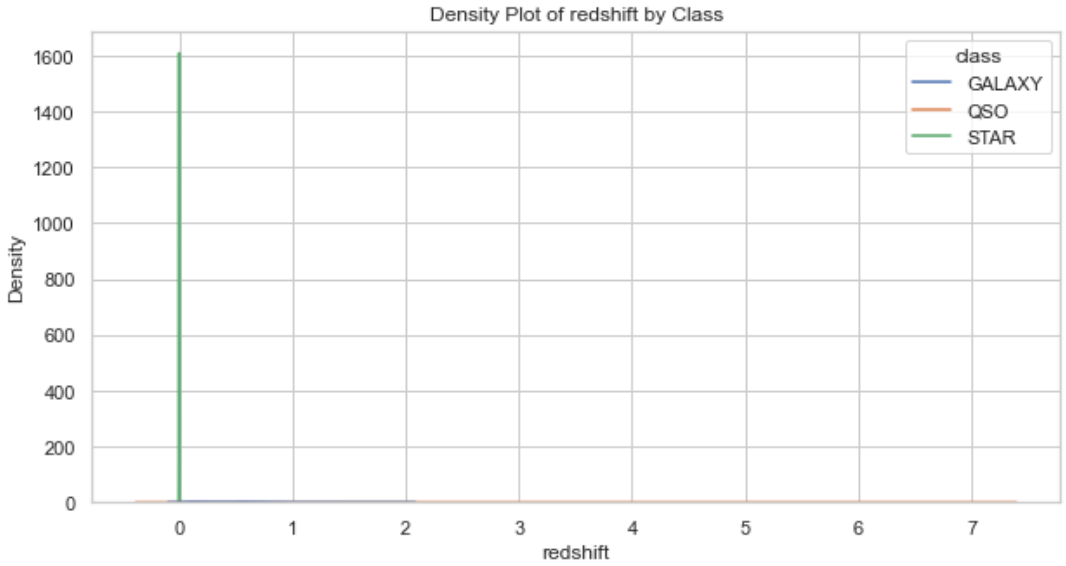

Figura 8: Densidad de la variable redshift.

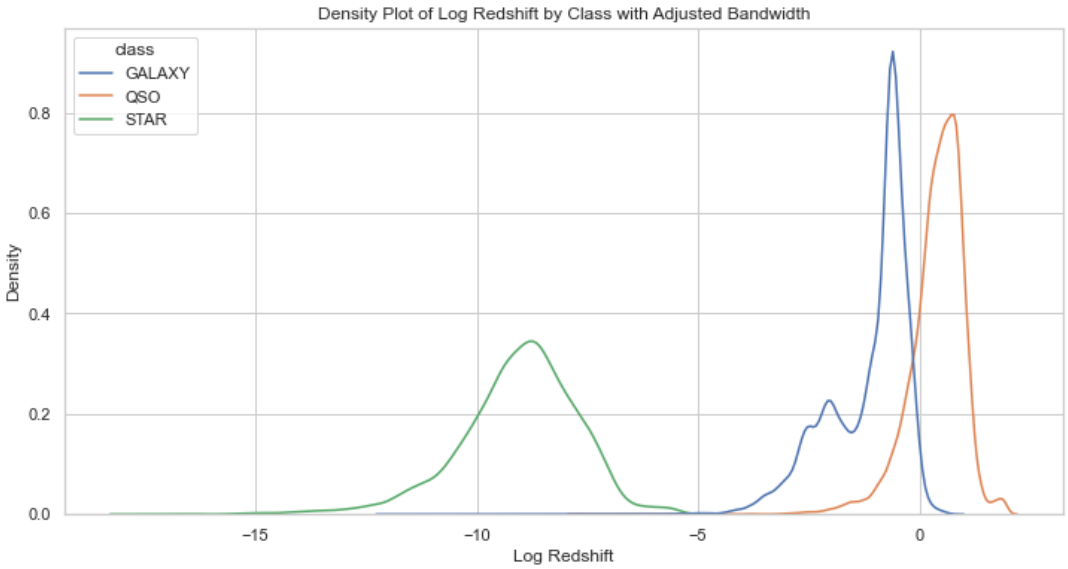

Figura 9: Transformación logarítmica de la variable redshift.

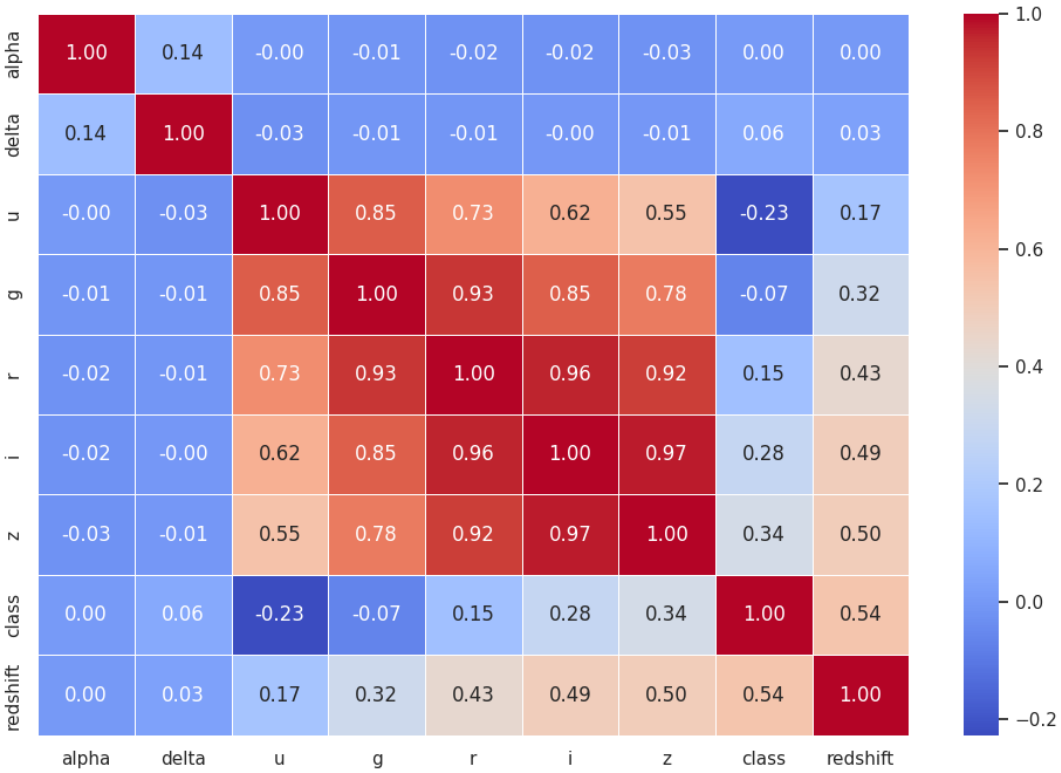

Figura 10: Matriz de correlación de Pearson.

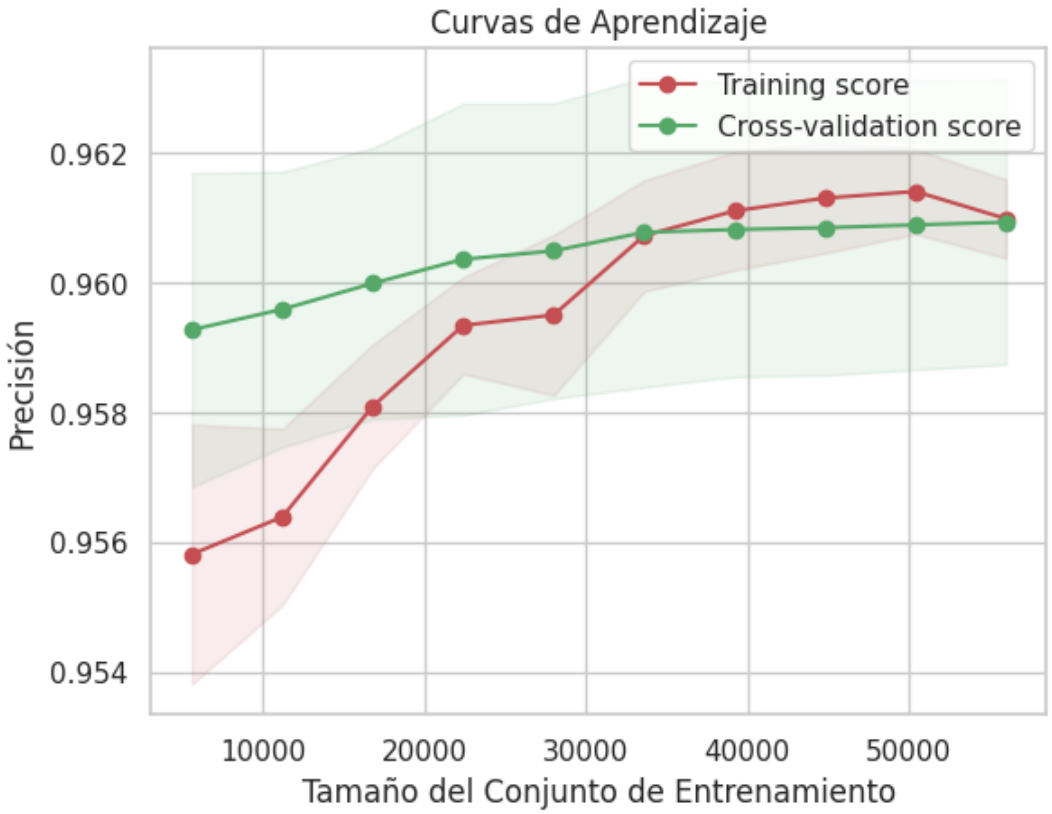

Figura 11: Curvas de aprendizaje

Referencias

[1] F. Z. Zeraatgari, F. Hafezianzadeh, Y. Zhang, L. Mei, A. Ayubinia, A. Mosallanezhad, and J. Zhang, "Machine learning-based photometric classification of galaxies, quasars, emission-line galaxies, and stars," *Monthly Notices of the Royal Astronomical Society*, vol. 527, no. 3, pp. 4677–4689, 2024. [Online]. Available: https://doi.org/10.1093/mnras/stad3436

[2] "Sloan Digital Sky Survey, SDSS Data Release 17," https://skyserver.sdss.org/dr17/en/home.aspx, 2023, accedido: 3/12/2023.

[3] "SDSS SkyServer DR17 SQL Search," https://skyserver.sdss.org/dr17/en/tools/search/sql.aspx, accedido: 3/12/2023.

[4] "L1 (Lasso) and L2 (Ridge) Regularizations in Logistic Regression," https://ai.plainenglish.io/l1-lasso-and-l2-ridge-regularizations-in-logistic-regression-53ab6c952f15, 2023, accedido: 4/12/2023.