UNIVERSIDAD DE LAS AMÉRICAS PUEBLA

ESCUELA DE CIENCIAS

DEPARTAMENTO DE ACTUARÍA, FÍSICA Y MATEMÁTICAS.

ESTIMACIÓN DE CORRIMIENTOS AL ROJO FOTOMÉTRICOS DE GALAXIAS Y CUÁSARES UTILIZANDO EXTRATREESREGRESSOR

TESIS QUE, PARA COMPLETAR LOS REQUISITOS DEL PROGRAMA DE HONORES PRESENTA LA ESTUDIANTE

SOFÍA ORDAZ LÓPEZ

167671

DIRECTOR

DRA. MILAGROS ZEBALLOS REBAZA

SAN ANDRÉS CHOLULA, PUEBLA. PRIMAVERA, 2024

DIRECTOR DE TESIS

Dra. Milagros Zeballos Rebaza

PRESIDENTE DE TESIS

Dra. Daniela Cortés Toto

SECRETARIO DE TESIS

Dr. Miguel Ángel Reyes Cortés

Índice

Agradecimientos

A mi ángel, Mónica Sevilla Pym.

A mi mamá Maximina, por ser mi gran ejemplo de superación y creatividad.

A mi papá Abel, por enseñarme que la dedicación y la disciplina son pilares esenciales para mi formación.

A mi hermanito Abel, por protegerme y ser mi mayor ejemplo de inteligencia y resiliencia.

A mi hermana Antonia, por ser mi mejor amiga, y mi mayor inspiración y mi fortaleza.

A mi Kalani, gracias por llegar a mi vida.

A la Dra. Milagros, por su mentoría e incansable motivación y dedicación.

Al Dr. Wanderson, por su tutoría y constante apoyo e inspiración.

A la Dra. Daniela y al Dr. Miguel, por las increíbles enseñanzas y asesorías.

A Alex, por su orientación, apoyo y amistad.

A Memo, Diana e Iñigo, por hacerme las clases y la vida más divertida.

A Marian y Wendy, por sus consejos y nunca dejarme sola.

A Paco, por ser mi gran compañero de aventuras.

A Tato, por su amor y cada día motivarme a ser mejor.

Resumen

El cálculo del corrimiento al rojo de galaxias y cuásares es fundamental en la cosmología moderna, ya que proporciona información crucial sobre la expansión del universo y la velocidad relativa de los objetos celestes. Dado que obtener observaciones espectroscópicas para una gran cantidad de galaxias puede ser difícil y costoso, se prefiere el uso de corrimientos al rojo fotométricos. El uso de técnicas de aprendizaje automático para estimar corrimientos al rojo fotométricos se ha convertido en un recurso valioso en la astronomía moderna debido a la gran cantidad de datos esperados de los próximos sensos astronómicos.

Actualmente existen algoritmos de aprendizaje automático que han demostrado poder estimar corrimientos al rojo fotométricos para valores de ($z < 1$). En este trabajo, se utiliza el algoritmo ExtraTreesRegressor de la librería ScikitLearn para obtener estimaciones en todo el rango de corrimiento al rojo ($0 < z < 7$) ofrecido por el catálogo del Sloan Digital Sky Survey (SDSS) mediante sus Data Releases del 12 al 18, utilizando métodos de preprocesamiento para limpieza de datos y selección de características de forma recursiva (RFE), un dominio logarítmico para evitar el desequilibrio de clases y un GridSearch para obtener los mejores hiperparámetros para el modelo, evaluándolo con las dos métricas más utilizadas en algoritmos de aprendizaje automático, Raiz del Error Cuadrático Medio (RMSE) y R^2.

Después de implementar recursos para optimizar el modelo de ExtraTreesRegressor, no fue posible obtener un MSE menor a 2.050, lo cual demuestra que el algoritmo no fue capaz de estimar con precisión los corrimientos al rojo en el rango ($0 < z < 7$). Sin embargo, al realizar el análisis para ($z < 1$), el MSE mejoró significativamente a 0.052. En conclusión, el algoritmo, con las características y parámetros detallados, solo funciona de manera efectiva para valores de ($z < 1$).

Palabras clave: ExtraTreesRegressor, corrimiento al rojo, galaxia, aprendizaje automático.

1. Introducción

El uso de técnicas de aprendizaje automático para estimar corrimientos al rojo ofrece diversas ventajas clave en comparación con enfoques tradicionales. Por ejemplo, entre las grandes ventajas están la capacidad de manejar y aprovechar los grandes volúmenes de datos generados por sondeos astronómicos modernos, la captura efectiva de patrones complejos, la habilidad de manejar características multidimensionales presentes en diferentes longitudes de onda, y la extracción automática de características relevantes para la estimación precisa de corrimientos al rojo. Además, el aprendizaje automático se adapta a cambios en los datos y se generaliza bien a nuevos conjuntos, lo que es crucial para la investigación astronómica. Además, permite estimaciones más eficientes y económicas en comparación con métodos costosos de observación espectroscópica. En conjunto, estas ventajas hacen que el aprendizaje automático sea una herramienta valiosa para mejorar la precisión y eficiencia en la estimación de corrimientos al rojo.

El telescopio SDSS ha llevado a cabo una serie de censos astronómicos a lo largo de sus diversas fases, cada una enfocada en objetivos específicos y la recopilación de datos esenciales para nuestra comprensión del universo. Dentro de estas fases, se destacan colaboraciones clave como: **SDSS Legacy Survey** que se completó en 2009 habiendo observado un cuarto del cielo en 5 filtros (**u-g-r-i-z**) y creado mapas tridimensionales conteniendo más de 190,000 galaxias y más de 120,000 cuásares, **eBOSS (Extended Baryon Oscillation Spectroscopic Survey)** que se enfoca en la recopilación de datos de corrimientos al rojo espectroscópicos de miles de millones de galaxias y cuásares distantes. Estos datos son cruciales para investigaciones sobre la expansión cósmica y la naturaleza de la energía oscura, **BOSS (Baryon Oscillation Spectroscopic Survey)** que se centró en la medición precisa de las oscilaciones bariónicas en la distribución de galaxias, lo que también contribuye a la comprensión de la expansión del universo y **SEGUE (Sloan Extension for Galactic Understanding and Exploration)** que se dedicó al estudio detallado de la estructura de la Vía Láctea, recopilando datos fotométricos y espectroscópicos de estrellas en nuestra galaxia (Alam y cols., 2015).

A medida que estos censos han progresado, la cantidad de datos generados ha crecido exponencialmente, para gestionar esta abrumadora cantidad de información, se ha recurrido al aprendizaje automático que se ha convertido en una herramienta invaluable en la astronomía moderna. El telescopio SDSS ha logrado obtener datos sobre galaxias y cuásares con corrimiento al rojo dentro del intervalo ($0 < z < 7$). Ya se han implementado técnicas de inteligencia artificial que buscan predecir los corrimientos al rojo de manera más precisa y eficiente, entre ellas podemos encontrar el uso de una red neuronal artificial para corrimientos al rojo (ANNz, por sus siglas en inglés), que resultó ser una herramienta competitiva y altamente precisa con un RMSE = 0.023 en el intervalo ($0 \leq z \leq 0.7$) con 50,000 muestras (Collister y Lahav, 2004), y en el mismo intervalo de corrimiento al rojo con un RMSE = 0.027 se implementó un algoritmo de máquinas de soporte vectorial para regresión (SVR, por sus siglas en inglés) pero esta vez con 139,000 muestras (Wadadekar, 2005). Ambos algoritmos demuestran ser muy precisos al estimar corrimientos al rojo a bajo rango; sin embargo, también se han realizado estudios para valores mayores a 1, por ejemplo, implementando una red neuronal convolucional (CNN, por sus siglas en inglés) combinada con una red de mixtura de densidades con resultados poco concluyentes (D'Isanto y Polsterer, 2018). Este trabajo de tesis tiene como enfoque estudiar paso a paso el trabajo realizado por Moonzarin Reza y Mohammad Ariful Haque titulado **Photometric redshift estimation using ExtraTreesRegressor: Galaxies and quasars from low to very high redshifts**, que promete poder estimar corrimientos al rojo de galaxias y cuásares en todo el rango de corrimiento al rojo que contiene el SDSS ($0 < z < 7$) con el algoritmo ExtraTreesRegressor con un Error Cuadrático Medio suficientemente bajo MSE = 0.66 (Reza y Haque, 2020).

1.1. Objetivo General

Estimar los corrimientos al rojo fotométricos de galaxias y cuásares utilizando el algoritmo de aprendizaje automático ExtraTreesRegressor.

1.1.1. Objetivos Específicos

- Comprender los conceptos y características principales de los objetos celestes a estudiar, así como la definición de corrimiento al rojo y la diferencia entre un corrimiento al rojo fotométrico y espectroscópico.

- Establecer la estructura del código SQL de manera adecuada para obtener los datos de la base de datos del SDSS.
- Implementar la función MinMaxScaler para normalizar los datos de entrada.
- Realizar la Eliminación Recursiva de Características, evaluando su impacto en el rendimiento del modelo y conservando solo aquellas que contribuyan significativamente a mejorar la precisión del algoritmo.
- Implementar el dominio logarítmico para contrarrestar el desequilibrio en la cantidad de datos disponibles a medida que aumenta el corrimiento al rojo.
- Utilizar un GridSearchCV para encontrar los mejores hiperparámetros para el modelo.
- Evaluar la eficacia del algoritmo realizando predicciones sobre el conjunto de prueba y obtener las métricas correspondientes.

1.2. Galaxias y cuásares

Previo al estudio y estimación de corrimientos al rojo fotométricos resulta esencial comprender las principales características de dos tipos de objetos astronómicos que se encuentran en las vastas bases de datos del Sloan Digital Sky Survey (SDSS): las galaxias y los cuásares.

1.2.1. Galaxias

Las galaxias son inmensas agrupaciones cósmicas que conforman los elementos fundamentales del universo, y varían notablemente en su composición y estructura. Algunas galaxias presentan una simplicidad sorprendente, compuestas principalmente por estrellas ordinarias y careciendo de características particulares notables. En contraste, otras galaxias son auténticas maravillas de complejidad, albergando una mezcla diversa de componentes, incluyendo estrellas, gas neutro e ionizado, polvo cósmico, nubes moleculares, campos magnéticos y rayos cósmicos. Además, las galaxias pueden agruparse en pequeñas comunidades o extenderse en enormes cúmulos en el vasto espacio del universo (ver Figura 1). En muchas ocasiones, el centro de una galaxia alberga un núcleo compacto que puede ser tan luminoso que supere en brillo a todas las radiaciones emitidas por el resto de la galaxia, a este núcleo se le suele denominar Núcleo Galáctico Activo o AGN (Active Galactic Nucleus, por sus siglas en inglés) (Karttunen, Kröger, Oja, Poutanen, y Donner, 2017).

La luminosidad de las galaxias es extraordinariamente diversa, oscilando desde aquellas que brillan con una intensidad miles de millones de veces mayor que la del Sol, hasta las más tenues que apenas alcanzan unas pocas veces la luminosidad solar. Cuantificar con precisión las masas y dimensiones de las galaxias puede ser un desafío, dado que estas estructuras cósmicas carecen de límites exteriores claramente definidos. En términos generales, una galaxia gigante puede acumular una masa que supera en miles de billones de veces la del Sol y extenderse a lo largo de varias decenas de miles de años luz en su radio. En contraste, las galaxias enanas son significativamente menos masivas y exhiben dimensiones más modestas. No obstante, es importante recordar que nuestra propia galaxia, la Vía Láctea, es solo una de las incontables galaxias que pueblan el universo. De hecho, la observación realizada por el Telescopio Espacial Hubble durante un lapso de 12 días reveló la presencia de alrededor de 10,000 galaxias de diversas formas, tamaños y colores en una pequeña porción del espacio. Este descubrimiento subraya la inmensidad del cosmos, y algunos científicos especulan que el número total de galaxias en el universo podría ascender a aproximadamente cien mil millones.

1.2.2. Cuásares

Un cuásar es un objeto astronómico extremadamente luminoso y distante que emite una gran cantidad de energía en forma de radiación electromagnética en diversas longitudes de onda, incluyendo radio, luz visible y rayos X. El descubrimiento de los cuásares tuvo lugar en 1963 gracias al trabajo del astrónomo holandés Maarten Schmidt. Schmidt analizó el espectro de una fuente de radio muy brillante conocida como 3C273 y encontró que las líneas espectrales de emisión estaban desplazadas hacia el rojo en un 16 %, indicando que esta fuente estaba alejándose de la Tierra a una velocidad extraordinaria y, por lo tanto, se encontraba a una inmensa distancia.

Este objeto, que no era ni una estrella ni una galaxia cercana, recibió el nombre de “cuásar”, abreviatura de ”fuente de radio cuasiestelar”. A pesar de que en las imágenes ópticas los cuásares parecen estrellas puntuales,

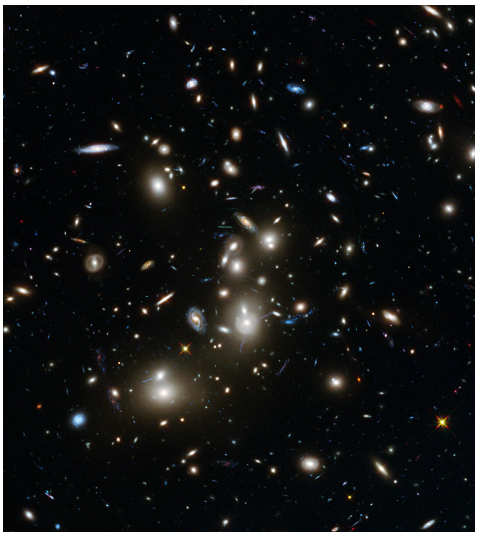

Figura 1: Cúmulo de galaxias Abell 2744 también conocido como cúmulo de Pandora. [Fotografía] por NASA, ESA y J. Lotz, M. Mountain, A. Koekemoer y el Equipo HFF (STScI), 2014, NASA/ESA (`https://esahubble.org/images/heic1401a/`).

en realidad son objetos astronómicos mucho más distantes y luminosos que cualquier estrella (Karttunen y cols., 2017).

Según el modelo unificado de núcleos activos de galaxias, los cuásares son agujeros negros súper masivos ubicados en el centro de galaxias, caracterizados por un disco de acreción compuesto de gas y polvo orbitando a su alrededor, con tal velocidad que este disco de material emite grandes cantidades de radiación electromagnética en un rango muy amplio de longitud de onda.

1.3. Efecto Doppler

En el estudio de la astrofísica y la astronomía el concepto de corrimiento al rojo o redshift resulta de vital importancia ya que se utiliza para medir la velocidad a la que los objetos astronómicos se alejan de nosotros y entre sí debido a la expansión del universo. Haciendo referencia a la expansión del universo y el corrimiento al rojo, vale la pena introducir el concepto de efecto Doppler.

El efecto Doppler es un fenómeno que se manifiesta como un cambio en la frecuencia de una onda debido al movimiento relativo entre la fuente y el receptor, no se limita a un tipo específico de onda, ya que puede observarse en diversas formas de ondas, incluyendo la luz y las ondas en el agua. Fue Gustav Doppler quien, en 1842, primero planteó esta teoría en su trabajo ”Sobre la luz coloreada de estrellas dobles y algunas otras estrellas del cielo”. La hipótesis encontró respaldo empírico cuando Buys Ballot, en 1845, confirmó que el tono de las ondas sonoras se elevaba al acercarse la fuente y disminuía al alejarse (Paik, 2021). Para definir la ecuación, se tomará como referencia el efecto Doppler en las ondas de sonido, exclusivamente en una situación especial en la que tanto la fuente como el receptor se desplazan a lo largo de una línea que los conecta. Primero se considera el caso en el que la fuente se encuentra en reposo mientras que el receptor se acerca o se aleja de la misma, entonces la frecuencia f' percibida por el receptor se define como:

$$f' = f(\frac{v \pm v_o}{v}) \tag{1}$$

en donde f es la frecuencia emitida por la fuente en reposo, v es la velocidad de la onda en el medio (como el aire), que se mantiene constante, y v_o es la velocidad del receptor en dirección a la fuente o alejándose de ella. Entonces la frecuencia percibida f' es igual a la frecuencia emitida f multiplicada por el cociente entre

v_o y v. El signo (+ o -) depende de si el receptor se acerca (+) y la frecuencia percibida aumenta o se aleja (-) y la frecuencia percibida disminuye. Ahora, si considerando la situación en la que el receptor se mantiene en reposo y es la fuente la que se encuentra en movimiento la frecuencia f' percibida por el receptor se define:

$$f' = f(\frac{v}{v \pm v_f}) \tag{2}$$

donde v_f es la velocidad a la que la fuente se aleja o acerca del receptor, donde el signo (+) se utiliza si la fuente se aleja del receptor y el signo (-) si la fuente se acerca al receptor. Finalmente, cuando tanto el receptor y la fuente se encuentran en movimiento es posible generalizar la ecuación de la frecuencia f' percibida por el receptor como:

$$f' = f(\frac{v \pm v_o}{v \pm v_f}) \tag{3}$$

Esta ecuación toma en cuenta tanto el efecto Doppler producido por el movimiento de la fuente como el efecto producido por el movimiento del receptor. El signo (+ o -) de las velocidades relativas se utiliza para indicar si la fuente y el receptor se están acercando o alejando entre sí, lo que afecta la percepción de la frecuencia del sonido (Robert Resnick, 1992).

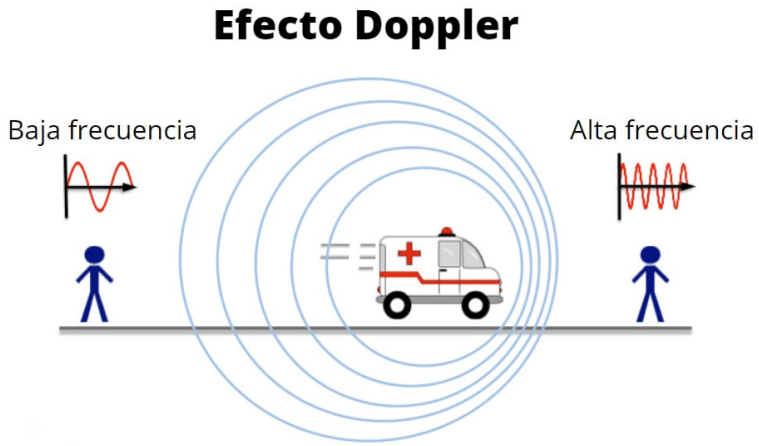

Figura 2: Efecto Doppler presente en la percepción del sonido de la sirena de una ambulancia. [Ilustración] por Fanny Zapata, 2020, Lifeder (`https://www.lifeder.com/efecto-doppler/`).

En la Figura 2 se ilustra la situación cuando la fuente de sonido se encuentra en movimiento (Ecuación 2). En este caso, se muestra como la frecuencia percibida por el observador cambia si la fuente se aleja (la longitud de onda aumenta, lo que resulta en una disminución de la frecuencia percibida) y cuando se aleja (la longitud de onda disminuye, lo que resulta en un aumento de la frecuencia percibida).

1.4. Corrimiento al rojo

En el contexto de las ondas de luz, se observa un fenómeno interesante conocido como ”desplazamiento al rojo” y ”desplazamiento al azul”, fenómenos que están directamente vinculados al efecto Doppler. En la Figura 3 se puede apreciar que, al disminuir la frecuencia y aumentar la longitud de onda de la luz, la desplaza hacia el extremo rojo del espectro electromagnético. Por el contrario, cuando la frecuencia aumenta y la longitud de onda disminuye, se produce un desplazamiento hacia el extremo azul.

Estos conceptos desempeñan un papel crucial en la astronomía, especialmente en nuestra comprensión del corrimiento al rojo cosmológico. A diferencia del desplazamiento al rojo y al azul, que están relacionados

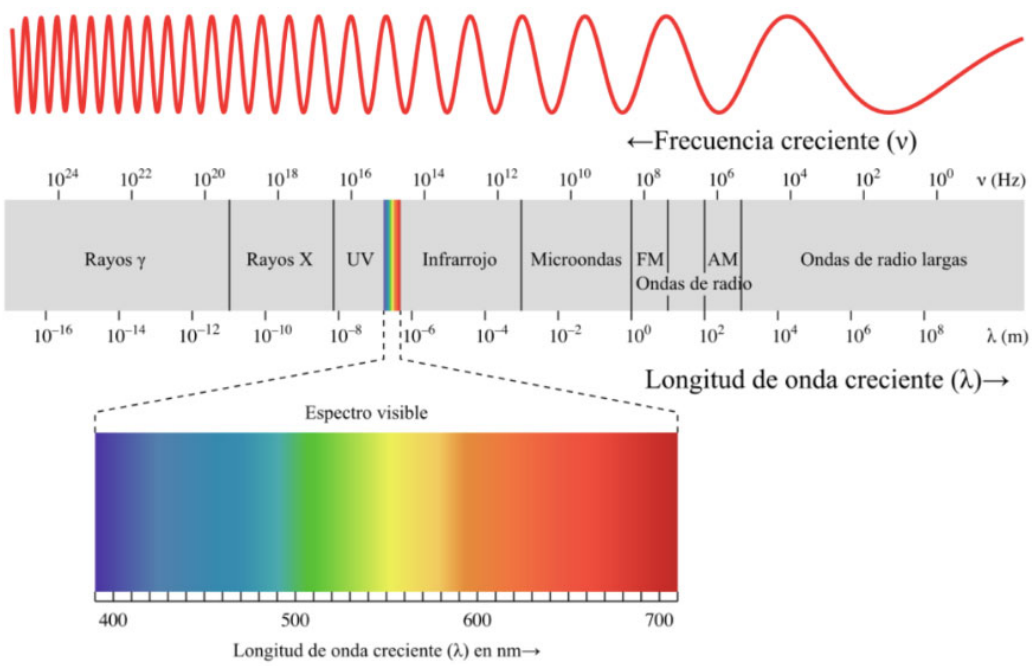

Figura 3: Espectro electromagnético [adaptación de Philip Ronan, Gringer (2008)] [Ilustración] por Bárbara Domínguez, Héctor Carecedo, Patricia Padilla y Josué García, 2020, Revista UNAM (`https://www.revista.unam.mx/2020v21n5/la_interaccion_de_la_luz_con_moleculas/`).

con el movimiento relativo de las fuentes de luz y sonido debido al efecto Doppler, el corrimiento al rojo cosmológico adquiere una perspectiva más amplia en el contexto de la expansión del universo.

La expansión cósmica provoca el estiramiento de las longitudes de onda de la luz, dando lugar al corrimiento al rojo cosmológico. Este fenómeno no solo es fascinante en sí mismo, sino que también se convierte en una herramienta esencial para medir la expansión cósmica y comprender la evolución histórica del universo a escalas cósmicas.

1.4.1. Ley de Hubble

Pero, ¿cómo llegamos a comprender la expansión del universo y su relación con el corrimiento al rojo? Edwin Hubble, pionero en la exploración de las nebulosas espirales, desempeñó un papel crucial en este descubrimiento. Durante sus estudios, Hubble notó un patrón intrigante en la luz emitida por estas nebulosas. Cuanto más alejadas estaban de la Tierra, más se desplazaba la luz hacia la parte roja del espectro. Este patrón condujo a una revelación asombrosa: las galaxias se alejaban de nosotros a velocidades proporcionales a su distancia. Esta relación, conocida como la ley de Hubble, establece que las galaxias más lejanas se alejan más rápidamente que las galaxias más cercanas. Es un principio fundamental que conecta el movimiento aparente de las galaxias con la expansión cósmica.

En el universo cercano la Ley de Hubble se expresa mediante la siguiente ecuación:

$$v = H_0 \times d \tag{4}$$

En donde v es la velocidad de recesión de las galaxias, H_0 es la constante de Hubble, que representa la tasa actual de expansión del universo, y d es la distancia a las galaxias.

Ahora, utilizando la ecuación del efecto Doppler cuando la fuente se mueve (Ecuación 2, podemos relacionar la velocidad de recesión v con el corrimiento al rojo z y la velocidad de la luz c. La relación es expresada por la fórmula:

$$z = \frac{\Delta\lambda}{\lambda_0} = \frac{v}{c} \tag{5}$$

donde z es el corrimiento al rojo cosmológico, $\Delta\lambda$ es el cambio en la longitud de onda observada, λ_0 es la longitud de onda emitida por la fuente, v es la velocidad de recesión de las galaxias, y c es la velocidad de la luz. Esta ecuación y la Ley de Hubble permiten conectar el corrimiento al rojo y la distancia a las galaxias en recesión, lo que nos permite comprender mejor la expansión del universo y la dinámica a gran escala de nuestro cosmos (Harrison, 2000).

1.5. Relevancia

La estimación precisa de corrimientos al rojo en objetos astronómicos desempeña un papel fundamental en nuestra comprensión del universo y su evolución. El corrimiento al rojo, derivado de la observación de desplazamientos en las longitudes de onda de la luz, proporciona información crucial sobre la velocidad y distancia a la que los objetos celestes se alejan de nosotros. Este fenómeno se relaciona directamente con la expansión del universo, permitiéndonos analizar la historia cósmica y comprender la distribución y movimiento de galaxias y cuásares en el espacio.

Existen dos métodos principales para determinar el corrimiento al rojo: espectroscópico y fotométrico. El primero implica analizar líneas espectrales específicas, mientras que el segundo se basa en mediciones de brillo en diferentes filtros. Aunque los métodos espectroscópicos son altamente precisos, su implementación a gran escala resulta costosa y laboriosa. Por el contrario, los datos fotométricos son más accesibles en términos de costo y esfuerzo, lo que los hace ideales para analizar grandes cantidades de objetos celestes, como millones de galaxias y cuásares en censos astronómicos.

La elección de la inteligencia artificial, en particular del algoritmo ExtraTreesRegressor basado en árboles de decisión, se justifica por su capacidad para manejar grandes volúmenes de datos y realizar estimaciones precisas. Este algoritmo, al utilizar múltiples árboles de decisión, puede capturar relaciones complejas en conjuntos de datos astronómicos, optimizando la predicción de corrimientos al rojo fotométricos. Las ventajas del ExtraTreesRegressor incluyen su robustez ante el sobreajuste, su capacidad para manejar datos heterogéneos y la eficiencia en la identificación de patrones no lineales (Geurts, Ernst, y Wehenkel, 2006). En el contexto de censos astronómicos en constante expansión, la aplicación de inteligencia artificial agiliza el proceso de análisis y permite una exploración más rápida y eficiente de la información del universo.

En resumen, este trabajo de tesis aborda la relevancia de estimar corrimientos al rojo de objetos celestes, la preferencia por datos fotométricos en censos astronómicos masivos y el papel crucial de la inteligencia artificial, en particular del algoritmo ExtraTreesRegressor.

2. Base de datos

Los datos utilizados para este estudio fueron extraídos del catálogo del Sloan Digital Sky Survey (SDSS) mediante los Data Releases 12, 13, 14, 15, 16, 17 y 18. Es importante destacar que entre estos lanzamientos existen datos repetidos debido a que los datos experimentan actualizaciones periódicas, con la inclusión de nuevas observaciones y la revisión de datos previamente registrados.

Para realizar el análisis, se crearon dos bases de datos distintas. Ambas bases de datos fueron aleatoriamente subdivididas, extrayendo 1,400 observaciones para el conjunto de prueba. Para garantizar un equilibrio en las clases, se seleccionaron 200 observaciones por cada intervalo de corrimiento al rojo. La primera base de datos contiene datos repetidos, abarcando los Data Releases 12, 13 y 14. En total, esta base de datos incluye 22,817 observaciones, de las cuales aproximadamente el 19 % son datos repetidos, la distribución de los datos

se puede observar en la Tabla 1, estas cuentas incluyen los 200 datos del conjunto de prueba por cada intervalo de corrimiento al rojo.

Corrimiento al rojo	Cantidad de datos
$[0-1)$	5200
$[1-2)$	5200
$[2-3)$	4472
$[3-4)$	2133
$[4-5)$	1925
$[5-6)$	1999
$[6-7)$	1888

Tabla 1: Cantidad de datos por intervalo de corrimiento al rojo.

La segunda base de datos se generó a partir de la primera, eliminando las observaciones repetidas y añadiendo datos de los Data Releases 15, 16, 17 y 18 para completar la cantidad original. En conjunto, esta base de datos final consta de 21,227 observaciones, la cantidad de datos por intervalo se detalla en la Tabla 2.

Corrimiento al rojo	Cantidad de datos
$[0-1)$	5200
$[1-2)$	5200
$[2-3)$	3494
$[3-4)$	1521
$[4-5)$	1925
$[5-6)$	1999
$[6-7)$	1888

Tabla 2: Cantidad de datos por intervalo de corrimiento al rojo.

2.1. Sloan Digital Sky Survey

El Sloan Digital Sky Survey (SDSS) se constituye como una colaboración científica internacional, que inició su travesía en el año 2000 con el objetivo de construir la imagen tridimensional más detallada del universo. Respaldado por la Alfred P. Sloan Foundation, la National Science Foundation y el Astrophysical Research Consortium, el SDSS ha desempeñado un papel clave en la expansión de nuestro entendimiento sobre la evolución cósmica a gran escala, desbloqueando conocimientos cruciales sobre la formación de estrellas y galaxias, la historia de la Vía Láctea, la naturaleza de los agujeros negros supermasivos y la ciencia detrás de la energía oscura. Ubicado en el Observatorio Apache Point en Nuevo México, el telescopio Sloan Foundation de 2.5 metros ha sido la herramienta esencial para realizar observaciones detalladas del espacio. Este telescopio específico utiliza un sistema fotométrico de cinco filtros **u-g-r-i-z** para capturar imágenes en el espectro visible y de corrimiento al rojo, generando espectros de millones de objetos astronómicos (Gunn y cols., 2006).

Dentro del SDSS, se desarrollan diversos censos, cada uno con un propósito distintivo:

- **SDSS Legacy Survey**
  Produjo un mapa tridimensional uniforme y calibrado de 1 millón de galaxias y 100,000 cuásares.
- **APOGEE-1 y APOGEE-2**
  Exploran la composición química y cinemática de estrellas en la Vía Láctea.
- **BOSS (Baryon Oscillation Spectroscopic Survey)**
  Investigación de oscilaciones bariónicas para entender la expansión del universo.
- **eBOSS (Extended Baryon Oscillation Spectroscopic Survey)**
  Continuación de BOSS para mapear la expansión cósmica.

- **MaNGA (Mapping Nearby Galaxies at Apache)**
  Se centra en la estructura y dinámica de galaxias cercanas.
- **MARVELS (Multi-Object APO Radial-Velocity Exoplanet Large-Area Survey)**
  Busca exoplanetas mediante mediciones de velocidad radial.
- **SEGUE (Sloan Extension for Galactic Understanding and Exploration)**
  Exploración de la Vía Láctea y búsqueda de supernovas.

Estos censos se lanzan y actualizan en distintas épocas, destacando SDSS-I, SDSS-II, SDSS-III, SDSS-IV, y la última incorporación, SDSS-V. Cada época presenta Data Releases, actualizaciones de datos que muestran los avances de los proyectos. Por ejemplo, SDSS-IV incluyó los proyectos APOGEE-2, MaNGA, y eBOSS, lanzando los Data Releases 13, 14, 15, 16 y 17 para actualizar los resultados de estos proyectos.

Un aspecto distintivo del SDSS es su compromiso con la accesibilidad de los datos. Los resultados están disponibles al público y se obtienen mediante un comando SQL en el SkyServer, fomentando la colaboración y la participación en la exploración del cosmos.

2.1.1. Data release

El SDSS actualiza sus datos de forma periódica, a estas actualizaciones se les conoce como **Data Releases**, con cada una de estas actualizaciones se publica también un artículo denominado **Data Release paper**, detallando la naturaleza de los datos, el proceso de adquisición y otros aspectos fundamentales del proyecto. En este proyecto, se han utilizado datos de los Data Releases 12, 13, 14, 15, 16, 17 y 18 (el más actual), cada uno contiene información sobre proyectos específicos, detallados a continuación:

- **Data Release 18**
  Disponible públicamente: 19 de enero de 2023
  Se enfoca en objetivos específicos y presenta los primeros espectros obtenidos de SDSS-V.
  Journal publication: `https://doi.org/10.3847/1538-4365/acda98`
- **Data Release 17**
  Disponible públicamente: 6 de diciembre de 2021
  Ofrece una visión completa de MaNGA, MaStar y los datos de APOGEE-2, abarcando diversos aspectos de la exploración.
  Journal publication: `https://doi.org/10.3847/1538-4365/ac4414`
- **Data Release 16**
  Disponible públicamente: 9 de diciembre de 2019
  Presenta los primeros resultados de la APOGEE-2 Southern Survey y el lanzamiento completo de los espectros de eBOSS.
  Journal publication: `https://doi.org/10.3847/1538-4365/ab929e`
- **Data Release 15**
  Disponible públicamente: 10 de diciembre de 2018
  Destaca por la liberación de las primeras cantidades derivadas de MaNGA, herramientas de visualización de datos y la biblioteca estelar.
  Journal publication: `https://doi.org/10.3847/1538-4365/aaf651`
- **Data Release 14**
  Disponible públicamente: 31 de julio de 2017
  Introduce los primeros datos espectroscópicos del extended Baryon Oscillation Spectroscopic Survey y de la segunda fase del Apache Point Observatory Galactic Evolution Experiment.
  Journal publication: `https://doi.org/10.3847/1538-4365/aa9e8a`
- **Data Release 13**
  Disponible públicamente: 31 de julio de 2016
  Presenta los primeros datos espectroscópicos de la SDSS-IV Survey Mapping Nearby Galaxies en Apache Point Observatory.
  Journal publication: `https://doi.org/10.3847/1538-4365/aa8992`
- **Data Release 11 y 12**
  Disponible públicamente: 6 de enero de 2015

Representan la culminación de SDSS-III, proporcionando los datos finales de esta fase del proyecto.
Journal publication: `https://doi.org/10.1088/0067-0049/219/1/12`

Cada Data Release representa un paso adelante en nuestra comprensión del cosmos, respaldado por la colaboración y dedicación de la comunidad científica involucrada en el SDSS. Para obtener información más detallada, se hace referencia a los artículos específicos de cada Data Release.

2.2. Filtros u - g - r - i - z

En el contexto del Sloan Digital Sky Survey (SDSS), la observación y clasificación de objetos astronómicos se lleva a cabo a través de filtros de color específicos, conocidos como los filtros U, G, R, I y Z. Estos filtros desempeñan un papel fundamental al permitir que se detecte la luz emitida por objetos celestes en diferentes longitudes de onda.

Cada filtro ha sido diseñado meticulosamente para dejar pasar la luz en un rango moderado de longitudes de onda centrado alrededor de una longitud de onda específica, bloqueando efectivamente la luz en todas las demás longitudes de onda. Este ancho de banda se encuentra descrito en la Tabla 3.

Filtro	Longitud de onda (Å)
u	3100 - 3900
g	3850 - 5400
r	5350 - 7000
i	6850 - 8500
z	8100 - 11000

Tabla 3: Rango de paso de banda de los cinco filtros.

La información recopilada a través de estos filtros proporciona datos valiosos sobre las propiedades espectrales de los objetos observados. A continuación, se detalla la función de cada filtro (Smith y cols., 2002):

- **Filtro U:** Este filtro permite la entrada de luz centrada alrededor de una longitud de onda más corta que el espectro visible. Al observar objetos a través del filtro U, se capturan las características asociadas a longitudes de onda ultravioletas.
- **Filtro G:** Diseñado para detectar luz verde en el espectro visible, el filtro G se centra en una longitud de onda ligeramente más larga que el filtro U. Esto abarca específicamente la región del verde-azul del espectro.
- **Filtro R:** Centrado en la región del espectro cercana al rojo, el filtro R permite el paso de luz en longitudes de onda correspondientes a esta región. Captura características específicas asociadas a estrellas y galaxias en esta parte del espectro.
- **Filtro I:** Al centrarse en la región infrarroja del espectro, el filtro I deja pasar la luz de longitudes de onda más largas que las visibles. Esto es esencial para estudiar objetos astronómicos que emiten predominantemente en el infrarrojo.
- **Filtro Z:** El filtro Z es el que deja pasar la luz emitida a longitudes de onda más largas que el filtro i, pero aún en la región del infrarrojo cercano. Permite la observación de objetos astronómicos con emisión significativa en estas longitudes de onda.

Los astrónomos utilizan estos filtros para recopilar datos que les permiten comprender mejor las características espectrales de estrellas, galaxias y otros objetos celestes.

En la Figura 4 se presentan las funciones de respuesta para cada filtro/detector en el conjunto fotométrico. Estas curvas muestran la eficiencia del sistema o QE (Quantum efficiency, por sus siglas en inglés), es decir, cuánta luz es detectada por el sistema en función de la longitud de onda de la luz incidente (λ).

Las curvas de respuesta incluyen la transmisión del filtro, la eficiencia cuántica (QE) para el detector de carga acoplada (CCD), la pérdida de flujo debido a los correctores ópticos y las reflectividades de dos

superficies de aluminio que también conforman el sistema. Además, se presentan las curvas de respuesta que incluyen la extinción atmosférica esperada con 1.2 masas de aire, lo que nos proporciona información sobre cómo la atmósfera afecta la eficiencia del sistema en diferentes longitudes de onda.

Esta información es crucial para comprender cómo el sistema responde a la luz en diferentes condiciones atmosféricas y cómo se ven afectadas las mediciones fotométricas. Los datos presentados en la Figura 4 nos ayudan a evaluar y comprender la eficiencia y la precisión del sistema en diferentes condiciones de observación.

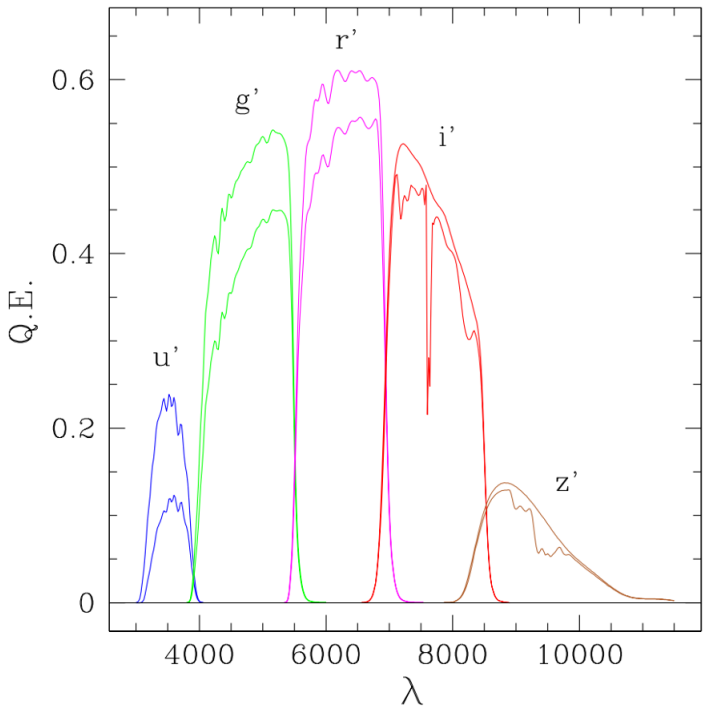

Figura 4: Eficiencia cuántica (QE) para cada sistema de filtro/detector en el conjunto fotométrico. El segundo conjunto de curvas muestra la respuesta del sistema incluyendo la extinción atmosférica esperada con 1.2 masas de aire. [Ilustración] por NASA, ESA y J. Lotz, M. Mountain, A. Koekemoer y el Equipo HFF (STScI), 2014, NASA/ESA (`https://esahubble.org/images/heic1401a/`).

2.3. Features

El enfoque principal de las bases de datos del SDSS se encuentra en las magnitudes en las cinco bandas **u-g-r-i-z**. Investigaciones previas basadas en algoritmos de aprendizaje automático han puesto su atención en las magnitudes de dichas bandas o en los colores definidos por la diferencia entre magnitudes de filtros adyacentes.

Con el objetivo de mejorar el rendimiento de los modelos estandares de aprendizaje automático, en este proyecto se busca replicar lo hecho por Moonzarin Reza y Mohammad Ariful Haque (Reza y Haque, 2020), incorporando más características adicionales y relevantes como características pertenecientes a clases considerablemente distintas. En total, hemos utilizado 30 características, abarcando magnitudes y colores de modelo, magnitudes de fibra, parámetros morfológicos, magnitudes de solapamiento de bandas y magnitudes promedio de filtros adyacentes (ver Tabla 4 a la 9).

2.3.1. Magnitudes y colores del modelo

Las magnitudes del modelo, representadas por modelMag_u, modelMag_g, modelMag_r, modelMag_i y modelMag_z (ver características 1-5 en la Tabla 4), indican las magnitudes aparentes en diferentes bandas del espectro electromagnético, específicamente en las bandas U, G, R, I y Z, respectivamente. Estas magnitudes son el resultado de ajustes de modelos de perfil de brillo, utilizando el método de mejor ajuste entre modelos de tipo DeV/Exp (De Vaucouleurs/Exponencial) para determinar la magnitud aparente en cada banda respectiva. Los datos se presentan en formato numérico real de 4 decimales.

Asímismo, las características 6, 7, 8 y 9 de la Tabla 4 hacen referencia a los colores del modelo. Estos

colores se obtienen calculando la resta entre las magnitudes en las bandas correspondientes. Por ejemplo, ModelColor_u-g se determina restando la magnitud en la banda G de la magnitud en la banda U. De manera similar, los otros colores se obtienen mediante la diferencia entre las magnitudes en las bandas adyacentes. Estos valores de color proporcionan información sobre las características espectrales y la distribución espectral de la luz emitida por los objetos astronómicos.

Índice	Característica
1	ModelMagnitude_u
2	ModelMagnitude_g
3	ModelMagnitude_r
4	ModelMagnitude_i
5	ModelMagnitude_z
6	ModelColor_u-g
7	ModelColor_g-r
8	ModelColor_r-i
9	ModelColor_i-z

Tabla 4: Índice y nombre de las primeras 9 características.

2.3.2. Magnitudes de fibra

Las magnitudes de fibra, por otro lado, hacen referencia al flujo lumínico de una fuente en un radio de 3 segundos de arco de diámetro utilizando una fibra óptica. Cada una de las magnitudes fiberMag_u, fiberMag_g, fiberMag_r, fiberMag_i y fiberMag_z corresponde al flujo lumínico en las bandas SDSS U, G, R, I y Z, respectivamente.

Al igual que las Magnitudes del modelo, los resultados de las Magnitudes de fibra se presentan en formato numérico real con cuatro decimales. Los colores se obtienen mediante la diferencia entre las magnitudes en las bandas adyacentes.

Índice	Característica
10	FiberMagnitude_u
11	FiberMagnitude_g
12	FiberMagnitude_r
13	FiberMagnitude_i
14	FiberMagnitude_z
15	FiberColor_u-g
16	FiberColor_g-r
17	FiberColor_r-i
18	FiberColor_i-z

Tabla 5: Índice y nombre de las Magnitudes de Fibra.

2.3.3. Parámetros morfológicos

En trabajos previos, la inclusión de parámetros morfológicos ha conducido a una mejora significativa en la precisión al estimar corrimientos al rojo (Gomes, Jarvis, Almosallam, y Roberts, 2017). En nuestro estudio, solo se han incluido dos características morfológicas: la relación del radio petrosiano que contiene el 50 % por ciento y el 90 % del flujo lumínico en la banda r y la relación correspondiente en la banda z (ver Tabla 6).

El radio petrosiano es un parámetro utilizado en la astronomía para caracterizar la extensión de la luz emitida por una galaxia. Se define como el radio a partir del cual la densidad superficial de flujo de la galaxia

cae por debajo de un valor específico en relación con la densidad superficial de flujo promedio del cielo circundante. En otras palabras, es el radio a través del cual se realiza una medición uniforme del flujo de luz de la galaxia, independientemente de su forma o brillo superficial. Estas medidas son fundamentales para comprender la morfología y la distribución de la luz en las galaxias, lo que a su vez puede proporcionar información valiosa sobre su evolución y formación (Reza y Haque, 2020).

Índice	Característica
19	Petroradius_{50}_r/Petroradius_{90}_r
20	Petroradius_{50}_z/Petroradius_{90}_z

Tabla 6: Parámetros morfológicos.

2.3.4. Magnitudes de superposición de bandas

Se puede observar en la Figura 4 que los filtros del telescopio SDSS exhiben una considerable superposición entre las bandas r-i e i-z. Para abordar esta superposición, se estimaron las magnitudes de las mediciones en la región de superposición de las bandas utilizando una interpolación lineal. Esto implicó calcular los valores de magnitud en los puntos donde las bandas se superponen mediante una función lineal que conecta los valores conocidos de las bandas adyacentes. Para ello se utilizó la notación mostrada en las Tablas 7 y 8 y las siguientes ecuaciones:

Interpolación r-i:

Banda	r	i
Longitud de onda	5350 - 7000	6850 - 8500
Flujo	a	b

Tabla 7: Interpolación lineal para los filtros r e i.

$$\text{Flujo } r = \frac{7000 - 6850}{7000 - 5350} \cdot a = \frac{a}{11} \tag{6}$$

$$\text{Flujo } i = \frac{7000 - 6850}{8500 - 6850} \cdot b = \frac{b}{11} \tag{7}$$

$$\text{Flujo total } ri = \frac{a}{11} + \frac{b}{11} \tag{8}$$

Interpolación i-z:

Banda	i	z
Longitud de onda	6850 - 8500	8100 - 11000
Flujo	a	b

Tabla 8: Interpolación lineal para los filtros i y z.

$$\text{Flujo } i = \frac{8500 - 8100}{8500 - 6850} \cdot a = \frac{8a}{33} \tag{9}$$

$$\text{Flujo } z = \frac{8500 - 8100}{11000 - 8100} \cdot b = \frac{4b}{29} \tag{10}$$

$$\text{Flujo total } iz = \frac{8a}{33} + \frac{4b}{29} \tag{11}$$

Este procedimiento resultó en la obtención de dos nuevas variables que representan a las características 21 y 22: ”Overlap between r-i band” y ”Overlap between i-z band”, las cuales contribuyen a eliminar errores

aleatorios debidos al ruido gaussiano y proporcionan estimaciones mejoradas de la relación señal-ruido (SNR) (Reza y Haque, 2020).

2.3.5. Magnitudes medias de filtros adyacentes

Se incluyeron ocho características calculando el promedio de las magnitudes de filtros adyacentes, tanto para las magnitudes de fibra como para las de modelo. Estas características se detallan en la Tabla 9 (Características 23 a 30) y se emplearon en el entrenamiento del modelo.

Índice	Característica
23	Adjacent_Mean_Model_u-g
24	Adjacent_Mean_Model_g-r
25	Adjacent_Mean_Model_r-i
26	Adjacent_Mean_Model_i-z
27	Adjacent_Mean_Fiber_u-g
28	Adjacent_Mean_Fiber_g-r
29	Adjacent_Mean_Fiber_r-i
30	Adjacent_Mean_Fiber_i-z

Tabla 9: Índice y nombre de las Magnitudes medias de filtros adyacentes.

2.4. Código SQL

Para extraer los datos utilizados en este proyecto, se accedió a un componente del telescopio SDSS denominado SkyServer. El SkyServer es una interfaz web que permite a los astrónomos y al público en general acceder a los datos recopilados por el SDSS y realizar consultas personalizadas sobre estos datos utilizando el lenguaje SQL.

Los data releases utilizados en este proyecto fueron el 12, 13, 14, 15, 16, 17 y 18. Los enlaces donde se pueden realizar las consultas SQL son los siguientes:

SDSS DATA RELEASE 12 `https://skyserver.sdss.org/dr12/en/tools/search/sql.aspx`

SDSS DATA RELEASE 13 `https://skyserver.sdss.org/dr13/en/tools/search/sql.aspx`

SDSS DATA RELEASE 14 `https://skyserver.sdss.org/dr14/en/tools/search/sql.aspx`

SDSS DATA RELEASE 15 `https://skyserver.sdss.org/dr15/en/tools/search/sql.aspx`

SDSS DATA RELEASE 16 `https://skyserver.sdss.org/dr16/en/tools/search/sql.aspx`

SDSS DATA RELEASE 17 `https://skyserver.sdss.org/dr17/en/tools/search/sql.aspx`

SDSS DATA RELEASE 18 `https://skyserver.sdss.org/dr18/en/tools/search/sql.aspx`

A continuación, se presenta la explicación del código SQL utilizado en este proyecto para extraer los datos necesarios.

```
SELECT TOP 5200
sp.objid as ID,
sp.ra as RA,
sp.dec as DEC,
```

Esta sección de código selecciona las primeras 5200 filas de un conjunto de datos astronómicos y extrae las columnas ”objid” (identificación del objeto astronómico), ”ra” (ascensión recta) y ”dec” (declinación), renombrándolas como ”ID”, ”RA” y ”DEC” respectivamente. La ascensión recta (RA) y la declinación (DEC) son coordenadas utilizadas en astronomía para ubicar objetos en el cielo. La ascensión recta es similar a la longitud en la Tierra y se mide en horas, minutos y segundos, desde el punto vernal o punto Aries hacia el este a lo largo del ecuador celeste. La declinación es similar a la latitud en la Tierra y se mide en grados, desde el ecuador celeste hacia los polos norte y sur.

Es por lo anterior que la ascensión recta y la declinación proporcionan una forma de localizar objetos en el cielo de manera precisa, lo que es fundamental para la astronomía y la cartografía celeste.

```
sp.modelmag_u as ModelMag_u,
sp.modelmag_g as ModelMag_g,
sp.modelmag_r as ModelMag_r,
sp.modelmag_i as ModelMag_i,
sp.modelmag_z as ModelMag_z,

sp.modelmag_u-sp.modelmag_g as ModelCol_ug,
sp.modelmag_g-sp.modelmag_r as ModelCol_gr,
sp.modelmag_r-sp.modelmag_i as ModelCol_ri,
sp.modelmag_i-sp.modelmag_z as ModelCol_iz,
```

Aquí se obtienen las magnitudes del modelo (Sección 2.3.1) y los colores, denominándolos como `ModelMag_u`, `ModelMag_g`, `ModelMag_r`, `ModelMag_i`, `ModelMag_z` para las magnitudes del modelo y `ModelCol_ug`, `ModelCol_gr`, `ModelCol_ri`, `ModelCol_iz` para los colores.

```
sp.fibermag_u as FiberMag_u,
sp.fibermag_g as FiberMag_g,
sp.fibermag_r as FiberMag_r,
sp.fibermag_i as FiberMag_i,
sp.fibermag_z as FiberMag_z,

sp.fibermag_u-sp.fibermag_g as FiberCol_ug,
sp.fibermag_g-sp.fibermag_r as FiberCol_gr,
sp.fibermag_r-sp.fibermag_i as FiberCol_ri,
sp.fibermag_i-sp.fibermag_z as FiberCol_iz,
```

Siguiendo la misma metodología que con las Magnitudes del modelo, se obtienen las Magnitudes de fibra (Sección 2.3.2) y sus correspondientes colores.

```
gx.petroR90_r/gx.petroR50_r as Petroradius90_r50_r,
gx.petroR90_z/gx.petroR50_z as Petroradius90_z50_z,
```

Este código calcula las razones entre dos medidas de radio petrosiano (Sección 2.3.3) en las bandas r y z. El cociente `Petroradius90_r/50_r` se calcula como el radio petrosiano a $90\,\%$ del flujo dividido por el radio petrosiano a $50\,\%$ del flujo en la banda r. De manera similar, `Petroradius90_z/50_z` se calcula para la banda z.

```
(sp.modelmag_r)/11 + (sp.modelmag_i)/11 as Overlap_ri,
(8*(sp.modelmag_i))/33 + (4*(sp.modelmag_z)/29) as Overlap_iz,
```

Para obtener las magnitudes de superposición de banda (Sección 2.3.4), se realizaron los cálculos mediante el código de búsqueda, utilizando la ecuación 8 para obtener `Overlap_ri` y la ecuación 11 para obtener `Overlap_iz`.

```
(sp.modelmag_u+sp.modelmag_g) / 2 as Adj_Mean_Model_ug,
(sp.modelmag_g+sp.modelmag_r) / 2 as Adj_Mean_Model_gr,
(sp.modelmag_r+sp.modelmag_i) / 2 as Adj_Mean_Model_ri,
(sp.modelmag_i+sp.modelmag_z) / 2 as Adj_Mean_Model_iz,

(sp.fibermag_u+sp.fibermag_g) / 2 as Adj_Mean_Fiber_ug,
(sp.fibermag_g+sp.fibermag_r) / 2 as Adj_Mean_Fiber_gr,
(sp.fibermag_r+sp.fibermag_i) / 2 as Adj_Mean_Fiber_ri,
(sp.fibermag_i+sp.fibermag_z) / 2 as Adj_Mean_Fiber_iz,
```

Para obtener la Magnitud media de los filtros adyacentes (Sección 2.3.5) el código de búsqueda calcula la media ajustada de las magnitudes de modelo y fibra para los filtros u, g, r, i y z. Para cada filtro, se suma la magnitud del filtro actual con la magnitud del filtro siguiente, y luego se divide entre 2 para obtener la media. Por ejemplo, `Adj_Mean_Model_ug` representa la media ajustada de las magnitudes de modelo para los filtros u y g, `Adj_Mean_Fiber_ri`representa la media ajustada de las magnitudes de fibra para los filtros r y i, y así sucesivamente.

```
sp.z as redshift,
sp.zErr as zErr,
sp.zwarning as zWarning,
sp.class as Class

FROM specPhoto AS sp
 JOIN galaxy AS gx ON sp.specobjid = gx.specobjid
 WHERE sp.z >= 0 and sp.z < 1 and sp.zErr <= 0.1 and (sp.zwarning = 0
 or sp.zwarning = 4) ORDER BY NEWID ()
```

Por último, se obtiene el corrimiento al rojo espectroscópico (`redshift`), el error asociado al corrimiento al rojo (`zErr`), la advertencia de corrimiento al rojo (`zWarning`) y la clase de objeto (`Class`) de la tabla `specPhoto`, uniéndola con la tabla `galaxy` mediante el campo `specobjid`. En el código ejemplo, se seleccionan únicamente las filas donde el corrimiento al rojo está entre 0 y 1, el error en el corrimiento al rojo es menor o igual a 0.1, y la advertencia de corrimiento al rojo es igual a 0 o igual a 4. Los acotamientos en el error de z y en las advertencias de z corresponden a datos tomados en las mejores condiciones. Es decir, siempre se seleccionan objetos cuyas observaciones tuvieron niveles aceptables de ruido. Los resultados se ordenan aleatoriamente para obtener una muestra variada.

3. Metodología

Para estimar los corrimientos al rojo fotométricos en todo el intervalo disponible por el SDSS ($0 < z < 7$), se empleó un enfoque de aprendizaje de máquina. Inicialmente, se utilizó un conjunto de 30 características relevantes para cada observación y se realizó una limpieza de datos que incluyó la eliminación de valores duplicados y la verificación de la existencia de datos nulos. Dado que el objetivo era predecir valores numéricos continuos, se optó por utilizar el algoritmo ExtraTreesRegressor, el cual se basa en árboles de decisión. Estos modelos dividen el conjunto de datos en subconjuntos más pequeños, basándose en las características de las observaciones, con el fin de realizar predicciones. La principal ventaja de ExtraTreesRegressor radica en su capacidad para reducir el sobreajuste y mejorar la precisión de las predicciones, especialmente en conjuntos de datos con un gran número de características, como el utilizado en este estudio.

Antes de aplicar ExtraTreesRegressor, se normalizaron las características utilizando MinMaxScaler para evitar sesgos en el modelo debido a diferencias en la escala. Además, se aplicaron técnicas comunes en los modelos de aprendizaje automático, como la evaluación de importancia de características, la eliminación recursiva de características, un dominio logarítmico para tratar el desequilibrio de clases en la cantidad de datos y un Grid Search, para mejorar el rendimiento del modelo y ajustar los hiperparámetros. A continuación se explican los conceptos más importantes.

3.1. Árboles de decisión

Los árboles de decisión son un tipo de modelo utilizado en aprendizaje automático que se basa en la estructura de un árbol para tomar decisiones. En este tipo de árbol, cada nodo interno representa una característica (o atributo), cada rama representa el resultado de una prueba en esa característica, y cada nodo hoja representa una clase o valor final.

En términos simples, un árbol de decisión toma una instancia y la pasa por el árbol de acuerdo con las pruebas en cada nodo. Comienza en el nodo raíz, que es el primer nodo del árbol y representa la característica que mejor divide el conjunto de datos inicial. Esta característica es elegida de acuerdo con algún criterio de partición, como la ganancia de información o la reducción de la impureza, que busca maximizar la homogeneidad de los subconjuntos resultantes. A medida que la instancia se mueve a través del árbol, cada nodo interno la guía hacia uno de sus hijos en función de la prueba realizada en ese nodo. Este proceso se repite hasta que la instancia llega a un nodo hoja, que proporciona la predicción final para esa instancia.

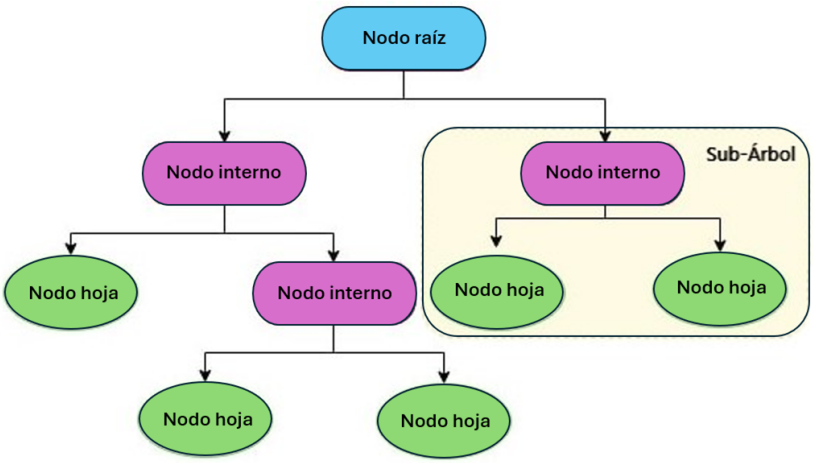

Figura 5: Diagrama de un árbol de decisión. [Ilustración] Adaptación por Sofía Ordaz de Chamanth Mvs, 2020, DataDrivenInvestor.

En el contexto de la regresión, los árboles de decisión se utilizan para predecir valores numéricos en

función de un conjunto de características. Cada hoja del árbol contiene un valor numérico que representa la predicción para las instancias que llegan a esa hoja (Rokach y Maimon, 2005). Durante el entrenamiento, el árbol de decisión se ajusta a los datos dividiendo el conjunto de datos en subconjuntos más pequeños en función de las características, de modo que las instancias en cada subconjunto sean lo más similares posible en términos de la variable objetivo (el valor que se está tratando de predecir).

3.1.1. ExtraTreesRegressor

ExtraTreesRegressor es un algoritmo de aprendizaje automático utilizado en problemas de regresión. Pertenece a la familia de algoritmos de aprendizaje por conjuntos, que combinan múltiples modelos para mejorar la precisión predictiva. En ExtraTrees, se construye un conjunto de árboles de decisión, donde cada árbol se entrena en una muestra aleatoria del conjunto de datos y se seleccionan aleatoriamente las características y los puntos de división para cada árbol. Esto se hace para reducir la varianza y mejorar la capacidad de generalización del modelo. La predicción final se obtiene promediando las salidas de todos los árboles en el conjunto (Geurts y cols., 2006).

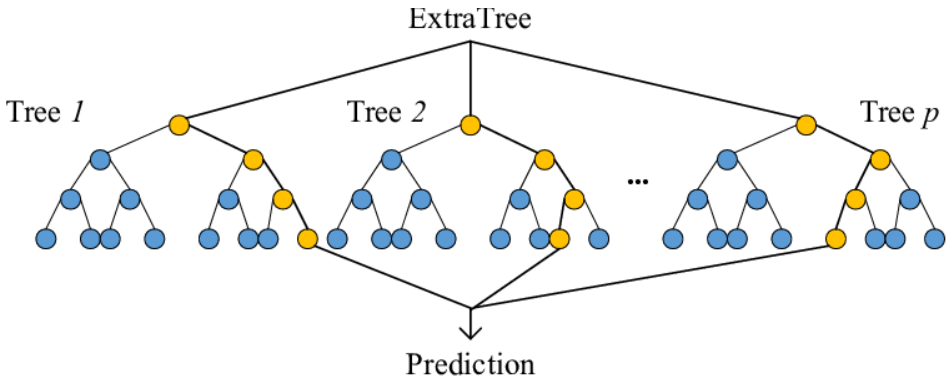

Figura 6: Estructura del algoritmo ExtraTrees [Ilustración] por Zheng Chu, Jiong Yu y Askar Hamdulla, 2020, School of Information Science and Engineering, Xinjiang University (https://doi.org/10.2298/CSIS200131031C).

Comparado con Random Forests, otro algoritmo de aprendizaje por conjuntos basado en árboles de decisión, ExtraTrees introduce una mayor aleatoriedad en la construcción de los árboles. Mientras que en Random Forests se busca el mejor umbral de división para cada característica en cada árbol, en ExtraTrees tanto las características como los puntos de división se eligen aleatoriamente. Esta mayor aleatoriedad en Extra Trees puede ayudar a prevenir el sobreajuste, ya que los árboles tienden a ser más diversos y menos propensos a memorizar el conjunto de datos de entrenamiento.

En el contexto de la estimación de corrimientos al rojo, ExtraTreesRegressor es una herramienta útil debido a su capacidad para manejar conjuntos de datos con muchas características y su capacidad para prevenir el sobreajuste. Además, su rapidez de entrenamiento puede ser una ventaja cuando se trabaja con grandes conjuntos de datos.

3.2. MinMaxScaler

MinMaxScaler es una función utilizada para transformar todos los valores dentro de una columna particular de un vector de características entre 0 y 1, mientras se conserva la forma de la distribución original. Esta técnica es útil cuando se trabaja con algoritmos de aprendizaje automático que son sensibles a la escala de las características, ya que ayuda a evitar que una característica domine sobre las demás debido a diferencias en la escala de sus valores.

En Python, para utilizar MinMaxScaler se debe importar desde la biblioteca sklearn.preprocessing de la

siguiente manera:

```
from sklearn.preprocessing import MinMaxScaler
```

Una vez importado, se debe crear una instancia de MinMaxScaler y utilizar el método fit_transform para aplicar la transformación a los datos. Por ejemplo:

```
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(data)
```

En este código, 'data' representa los datos originales, y 'scaled_data' será una versión de los datos donde los valores de cada característica han sido transformados para estar en el rango de 0 a 1.

Es importante mencionar que para el análisis realizado, todas las pruebas se hicieron con los datos normalizados utilizando MinMaxScaler.

3.3. Importancia de características

Inicialmente, se utilizaron las 30 características descritas en la sección 2.3 y se realizó un análisis de importancia de características. El comando "feature_importances_" del algoritmo "ExtraTreesRegressor" en Python proporciona una medida de la importancia de cada característica en el proceso de regresión. Esta medida se basa en la reducción de la impureza que cada característica aporta al modelo.

El criterio utilizado se conoce como "importancia Gini" y se calcula de la siguiente manera:

1. Para cada árbol en el conjunto de árboles entrenados, se mide cuánto disminuye la impureza total (normalizada) cuando se divide un nodo utilizando una característica específica.
2. Se promedian estos valores de reducción de impureza sobre todos los árboles para obtener la importancia de esa característica en el modelo general.
3. Finalmente, se normalizan estos valores para que sumen 1, a menos que todos los árboles sean árboles de un solo nodo que consistan solo en el nodo raíz, en cuyo caso la importancia de todas las características será 0.

La "impureza" se refiere a la medida de qué tan mezcladas están las clases en un nodo dado del árbol. En otras palabras, indica cuánto desorden o incertidumbre hay en los datos en ese nodo en términos de la variable objetivo que se está tratando de predecir. Hay diferentes medidas de impureza que se pueden utilizar, y una de las más comunes es el criterio de Gini. El criterio de Gini mide la probabilidad de que un elemento seleccionado al azar de un nodo dado sea clasificado incorrectamente si se le asigna una etiqueta al azar de acuerdo con la distribución de las etiquetas en el nodo.

Por lo tanto, la reducción en la impureza total al dividir un nodo utilizando una característica específica se refiere a cuánto disminuye la incertidumbre en la clasificación de los datos al dividir el nodo en subconjuntos más homogéneos en términos de la variable objetivo.

El comando se puede utilizar de la siguiente manera en Python:

```
importancia = modelo.feature_importances_
```

3.4. Eliminación recursiva de características

Después de obtener la importancia de cada característica, se procede a la selección de características, un paso crucial en el proceso de modelado de datos que determina qué características son las más relevantes para el algoritmo de aprendizaje automático. En este contexto, se utiliza la función Recursive Feature Elimination (RFE) en Python, la cual elimina características de manera recursiva evaluando el impacto de cada una en el rendimiento del modelo y conservando solo aquellas que contribuyen significativamente a mejorar la precisión del algoritmo.

El proceso de RFE comienza considerando todas las características disponibles y, de forma iterativa, elimina las menos importantes según un criterio definido. Este criterio se basa en la importancia de las características, medida a través de métodos como feature_importances_ en Python, que proporciona una puntuación para cada característica. Al eliminar las características menos importantes, se busca mantener solo aquellas que realmente aportan información relevante al modelo. El comando utilizado es el siguiente:

```
rfe = RFE(estimator=modelo)
```

Utilizar RFE en lugar de eliminar características manualmente tiene varias ventajas. En primer lugar, RFE es un proceso automático que puede aplicarse a conjuntos de datos grandes y complejos de manera eficiente. Además, al basarse en la importancia de las características, RFE ayuda a evitar sesgos y a seleccionar un conjunto óptimo de características que maximizan el rendimiento del modelo.

3.5. Dominio logarítmico

En la Tabla 2 se observa cómo la cantidad de datos disponibles disminuye a medida que aumenta el corrimiento al rojo. Para contrarrestar este desequilibrio, se emplea un enfoque de dominio logarítmico, dado que este desequilibrio puede impactar la precisión de los modelos de aprendizaje automático.

Se utiliza una versión adaptada del algoritmo de agrupamiento intraclase, efectivo en conjuntos de datos asimétricos, para el análisis de regresión. En este método, se aplica una transformación logarítmica a los valores reales de corrimiento al rojo, definida como:

$$z' = c \cdot \log_a(1 + z)$$

Aquí, 'c' y 'a' son constantes, con 'c' igual a 1 y 'a' igual a e (logaritmo natural). 'z' representa el corrimiento al rojo real, y 'z' ' es el corrimiento al rojo en el dominio logarítmico.

El algoritmo se entrena con datos en el dominio logarítmico, lo que hace que las predicciones también estén en esta misma escala. Finalmente, para obtener los valores de corrimiento al rojo predichos en su escala original, se realiza una operación exponencial dada por:

$$z = e^{z'} - 1$$

3.6. GridSearch

En un algoritmo como ExtraTreesRegressor, los hiperparámetros son configuraciones ajustables que no se aprenden durante el entrenamiento del modelo, sino que se establecen antes de iniciar el proceso de entrenamiento. Estos hiperparámetros afectan directamente al rendimiento y la capacidad del modelo para ser generalizado y utilizado con nuevos datos.

Por ejemplo, en ExtraTreesRegressor, los hiperparámetros como 'n_estimators' (número de árboles en el bosque), 'max_depth' (profundidad máxima de cada árbol), 'min_samples_split' (número mínimo de muestras requeridas para dividir un nodo interno) y 'min_samples_leaf' (número mínimo de muestras requeridas en un nodo hoja) influyen en la complejidad y la capacidad de generalización del modelo. Encontrar los mejores valores para estos hiperparámetros es crucial, ya que puede llevar a un modelo más preciso y con mejor capacidad de generalización.

El Grid Search es una técnica utilizada para encontrar los mejores hiperparámetros para un modelo. Consiste en definir una cuadrícula de posibles valores para cada hiperparámetro que se desea ajustar, y luego evaluar el rendimiento del modelo para cada combinación de valores de hiperparámetros en la cuadrícula. Esto se hace mediante validación cruzada, donde se divide el conjunto de datos en varios subconjuntos para entrenar y evaluar el modelo de forma iterativa. La cuadrícula de valores utilizada fue la siguiente:

```
param_grid = {'n_estimators': [100, 200, 300],
      'max_depth': [None, 5, 10, 15],
      'min_samples_split': [2, 5, 10],
      'min_samples_leaf': [1, 2, 4]}
```

Esta cuadrícula fue configurada así por las siguientes razones:

- **n_estimators:** Se eligió un rango típico de valores para el número de árboles en el bosque aleatorio. Usualmente, se comienza con un número moderado y se incrementa gradualmente. En este caso, se prueba con 100, 200 y 300 para cubrir una variedad de opciones y ver cómo afecta al modelo el número de árboles en el bosque.
- **max_depth:** La profundidad máxima de un árbol de decisión es crucial para controlar la complejidad del modelo y evitar el sobreajuste. Incluir None permite que el árbol crezca hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos ejemplos que min_samples_split. Luego, se prueban varias profundidades máximas para observar cómo afecta al rendimiento del modelo.
- **min_samples_split:** Este parámetro controla el número mínimo de muestras necesarias para dividir un nodo interno. Valores bajos pueden llevar a un sobreajuste, mientras que valores altos pueden conducir a un sesgo. Por lo tanto, se eligieron valores típicos para explorar cómo afecta al rendimiento del modelo.
- **min_samples_leaf:** Similar a min_samples_split, min_samples_leaf controla el número mínimo de muestras necesarias para estar en un nodo hoja. Valores bajos pueden causar sobreajuste, mientras que valores altos pueden causar subajuste. Al igual que con los otros parámetros, se eligieron valores comunes para observar cómo afecta al rendimiento del modelo.

3.7. Bootstrap

El Bootstrap, también conocido como bootstrapping, es una técnica de remuestreo que tiene como objetivo aproximar la distribución muestral de un estadístico. Se utiliza con frecuencia para estimar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza o realizar pruebas de hipótesis sobre parámetros de interés. En el contexto de esta investigación, se implementó Bootstrap para obtener intervalos de confianza alrededor de las predicciones de corrimiento al rojo realizadas por el algoritmo Extra Trees Regressor.

La lógica detrás de esta implementación consistió en lo siguiente:

- Tomar un conjunto de entrenamiento y seleccionar aleatoriamente 200 datos para formar el conjunto de prueba. Las predicciones se realizaron sobre este mismo conjunto de prueba en todas las repeticiones.
- Luego, realizar 1000 repeticiones Bootstrap, donde cada vez se tomó una muestra aleatoria del conjunto de entrenamiento (sin reemplazo) para entrenar el modelo ExtraTreesRegressor y realizar predicciones en el conjunto de prueba fijo.
- Para cada repetición, calcular el RMSE y R^2 entre las predicciones y los valores reales del conjunto de prueba.
- Promediar los valores de RMSE y R^2 obtenidos en las 1000 repeticiones para obtener un único valor representativo.
- Calcular los intervalos de confianza al 95 % para posteriormente graficar las bandas de confianza.

La realización de este análisis Bootstrap puede proporcionar una evaluación más robusta de la precisión de las predicciones del modelo. Al considerar la variabilidad inherente en los datos de entrenamiento y la sensibilidad del modelo a diferentes subconjuntos de datos de entrenamiento, se pueden obtener estimaciones más precisas y confiables de las predicciones. Esto permite una mejor interpretación de los resultados obtenidos y una mayor confianza en la validez de las conclusiones extraídas.

4. Análisis y Resultados

En esta sección se presentan y detallan los resultados de la implementación de cada paso descrito previamente en la sección de metodología. Se comenzará mostrando los resultados de los corrimientos al rojo estimados por el algoritmo Extra Trees Regressor, tanto con datos repetidos como después de eliminarlos, en todo el intervalo ($0 < z < 7$). Luego, se llevará a cabo un análisis para cada intervalo de corrimiento al rojo para observar el rendimiento individual. Posteriormente, se analizará la importancia de las características y la eliminación recursiva de las mismas. Además, se incluirá el análisis del dominio logarítmico y los resultados del Grid Search. Finalmente, se presentarán los intervalos de confianza obtenidos mediante Bootstrap para evaluar la variabilidad de las predicciones y se discutirán las métricas de evaluación, como el RMSE y R^2 promedio, para evaluar la precisión y la capacidad predictiva del modelo en la estimación de los corrimientos al rojo fotométricos.

El código de la implementación completa se encuentra en el siguiente enlace: `https://colab.research.google.com/drive/1PYnwk_l7b8T0v9BNi8bSMHK7s_ff7VTu?usp=sharing`

4.1. Primeras pruebas

Para las primeras pruebas se empleó el conjunto completo de datos disponible en el intervalo ($0 < z < 7$).

4.1.1. Con repetición

Uno de los principales objetivos era resaltar la diferencia que supone la presencia de datos duplicados en el conjunto de observaciones, ya que en el artículo (Reza y Haque, 2020) al parecer utilizaron datos repetidos. Esta primera prueba incluyó un 19 % de datos repetidos. Los datos utilizados en esta prueba se detallan en la Sección 2, Tabla 1.

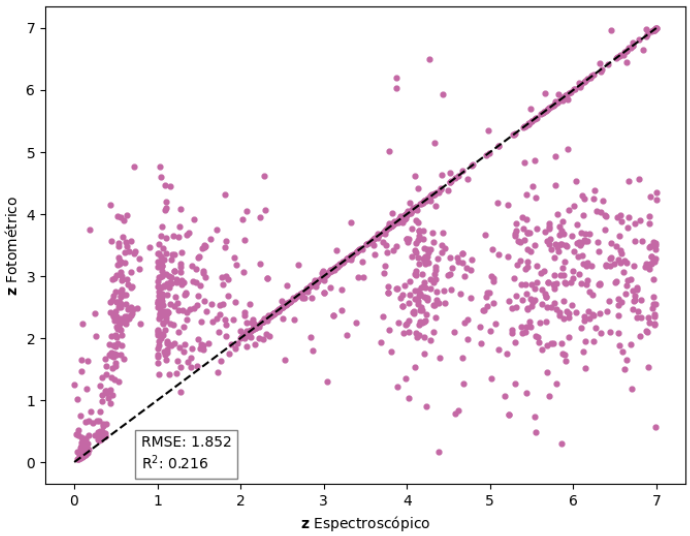

Figura 7: Corrimientos al rojo fotométricos obtenidos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos para el conjunto de galaxias y cuásares del SDSS. Estos resultados incluyen un 19 % de datos repetidos.

La Figura 7 muestra los corrimientos al rojo fotométricos obtenidos con ExtraTreesRegressor utilizando datos repetidos entre ($0 < z < 7$) contra los corrimientos al rojo espectroscópicos, es decir, los corrimientos al rojo reales.

Aunque la gráfica muestra cierta dispersión, es notable que algunos datos se ajustan a la función identidad (línea punteada), lo que sugiere algunas predicciones precisas. Sin embargo, tanto el RMSE como el R^2 no respaldan la conclusión de una estimación precisa generalizada.

4.1.2. Sin repetición

Se utilizó la base de datos de la prueba anterior, pero eliminando los datos duplicados y agregando nuevos datos de otros Data Releases para alcanzar la cantidad original. Con estos ajustes, se obtuvo la gráfica mostrada en la Figura 8.

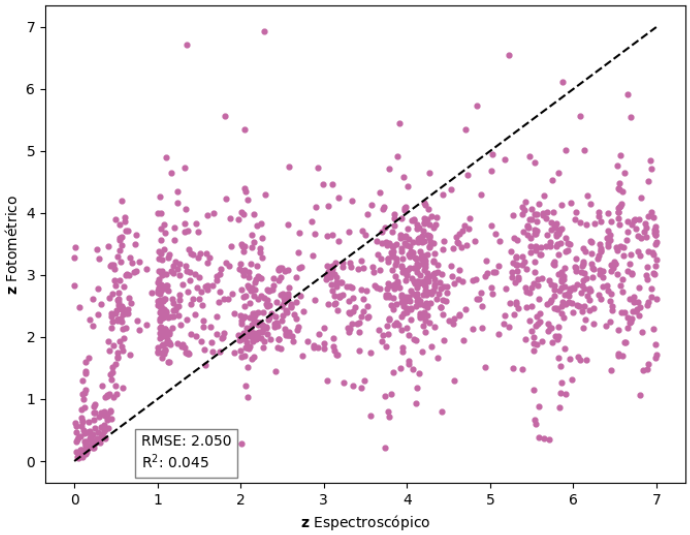

Figura 8: Corrimientos al rojo fotométricos obtenidos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos para las galaxias y cuásares del SDSS en el rango ($0 < z < 7$) después de realizar la limpieza de datos repetidos.

Gráficamente, se observa que ya no hay una correlación positiva evidente. Los datos parecen estar dispersos y las métricas indican un bajo rendimiento del modelo. Sin embargo, un análisis más detallado de la gráfica revela que podría existir una correlación en el intervalo de z entre 0 y 0.8.

Tanto la prueba ”con repetición” y ”sin repetición” fueron implementadas para intentar replicar la gráfica del resultado obtenido en el artículo **Photometric redshift estimation using ExtraTreesRegressor: Galaxies and quasars from low to very high redshifts**. Esta gráfica se muestra en la Figura 9. Este artículo reportó obtener un MSE = 0.66.

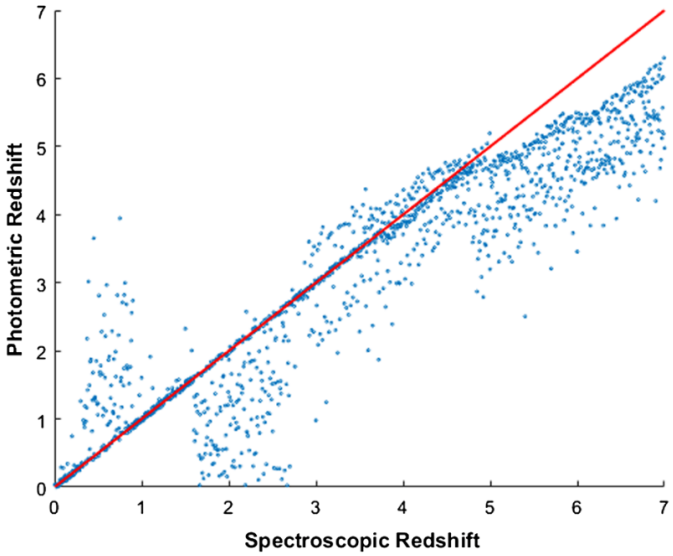

Figura 9: Comparación entre corrimientos al rojo fotométricos obtenidos en el artículo (Reza y Haque, 2020) y sus correspondientes valores espectroscópicos. Los objetos son galaxias y cuásares del SDSS con ($0 < z < 7$). Esta gráfica corresponde a la Figura 55 del artículo mencionado.

Como se puede apreciar en las Figuras 7, 8 y 9, los resultados reportados en (Reza y Haque, 2020) no fueron posibles de replicar. Además, dichas gráficas sugieren que el grado de éxito del artículo podría explicarse mediante la ausencia de la limpieza de datos repetidos.

4.1.3. Pruebas por intervalo

Basándonos en la conclusión de la última prueba, consideramos pertinente realizar una prueba individual para cada intervalo de corrimiento al rojo, por ejemplo, z: 0 - 1, z: 1 - 2, z: 2 - 3, y así sucesivamente. La cantidad de datos por intervalo se detalla en la Tabla 2, y se considera que para cada intervalo se selecciona un conjunto de prueba aleatorio de 200 datos. Los resultados se presentan en las Figuras 10, 11 y en el Anexo 6.1 (resto de intervalos).

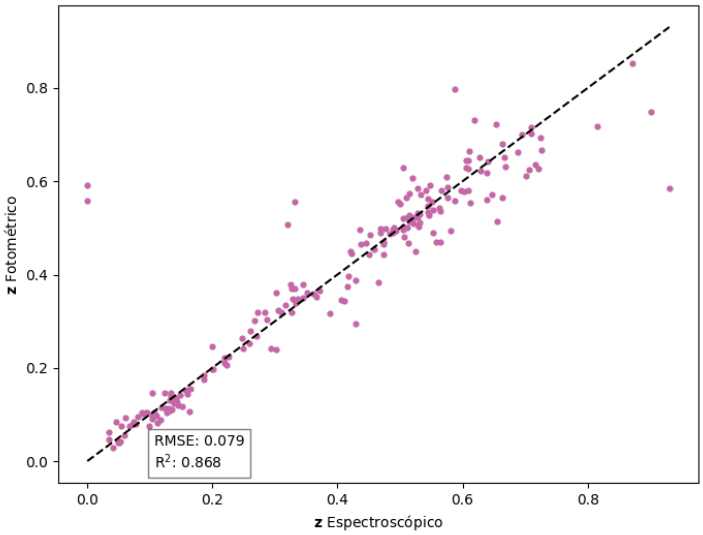

Figura 10: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($0 < z < 1$).

Como se puede apreciar en la Figura 10, la prueba para el primer intervalo, z: 0 - 1, ahora muestra una tendencia cercana a la función identidad. Con un RMSE de 0.079 y un R^2 de 0.868, se evidencia un buen rendimiento del modelo.

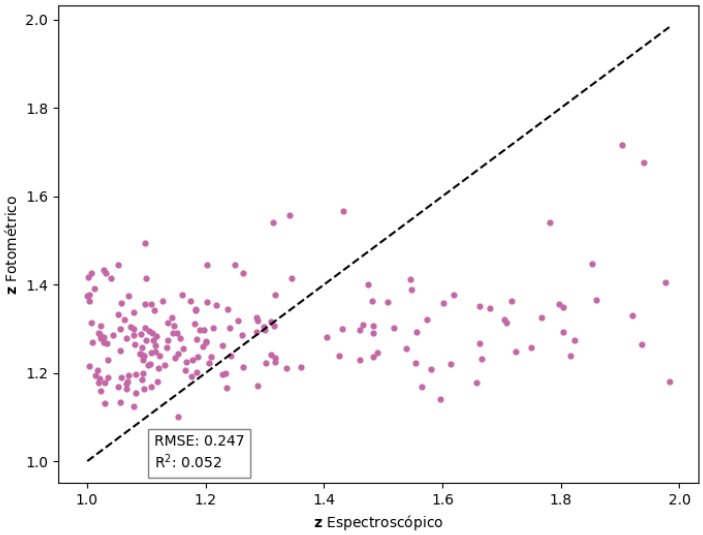

Figura 11: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($1 < z < 2$).

Por otro lado, la Figura 11 no muestra ninguna correlación aparente entre los datos estimados y los datos

reales. Además, el RMSE aumentó y el coeficiente de correlación R^2 disminuyó. Las pruebas para todos los demás intervalos son similares a esta, y las gráficas se encuentran en el Anexo 6.1.

4.2. Importancia de características

Dado el resultado obtenido en las pruebas individuales, todo el estudio se llevará a cabo de manera similar, es decir, todo se realizará por intervalos de corrimiento al rojo. El siguiente análisis nos proporcionará información sobre cómo el comando ”feature_importances_” logra distinguir, clasificar u otorgar importancia a cada característica del modelo.

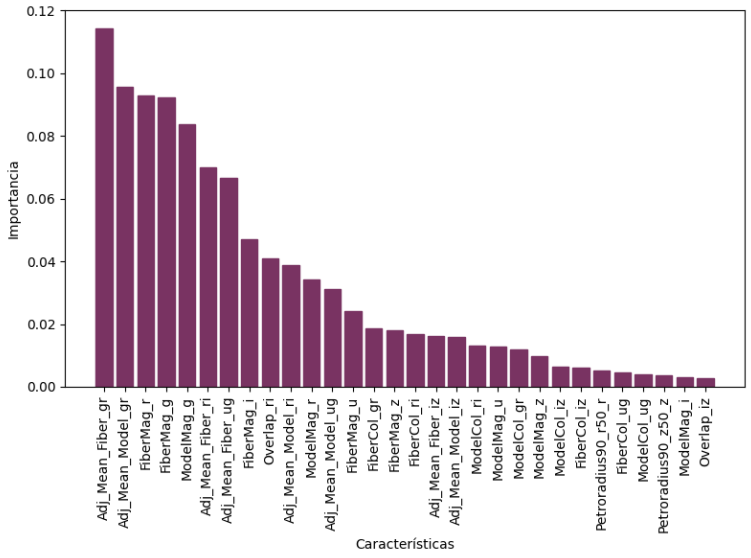

Figura 12: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($0 < z < 1$).

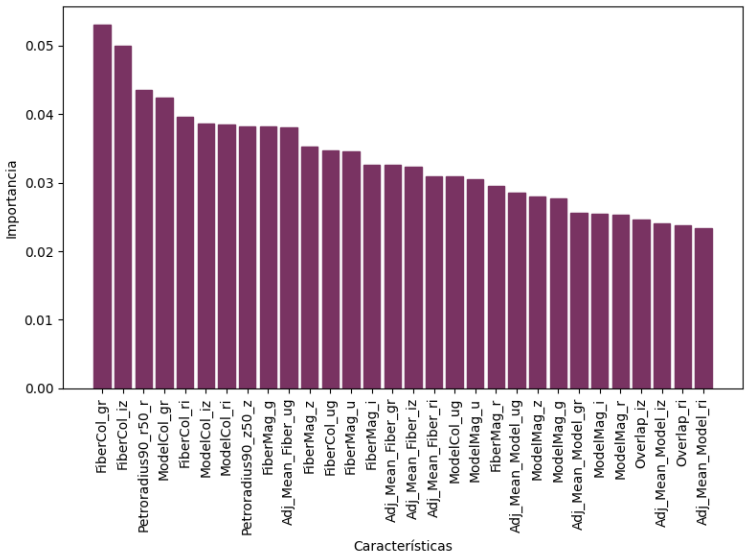

Figura 13: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($1 < z < 2$).

La Figura 12 muestra los resultados de importancia para cada una de las 30 características utilizadas en este trabajo y descritas en la Sección 2.3. Esta figura muestra resultados obtenidos con galaxias y cuásares con corrimientos al rojo entre 0 y 1. Como se aprecia, la gráfica muestra claramente que el algoritmo logra distinguir cómo algunas características contribuyen de manera más o menos significativa al modelo.

Por otro lado, la Figura 13 muestra la importancia obtenida para las mismas 30 características en el intervalo ($1 < z < 2$). Como se puede apreciar, ahora esta diferencia no es tan evidente, ya que a todas las características se les asigna prácticamente la misma importancia o una muy similar.

Los resultados para los demás intervalos de corrimiento al rojo se muestran en el Anexo 6.2, donde se observará el mismo resultado que para el intervalo ($1 < z < 2$).

4.3. Eliminación recursiva de características

En esta sección se presenta el resultado de la eliminación recursiva de características descrita en la Sección 3.4, es decir, las características seleccionadas para realizar las estimaciones de corrimiento al rojo.

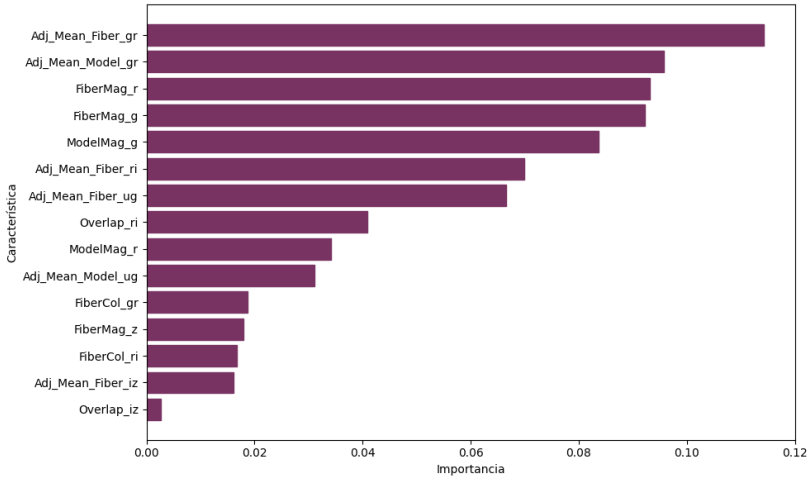

Figura 14: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($0 < z < 1$).

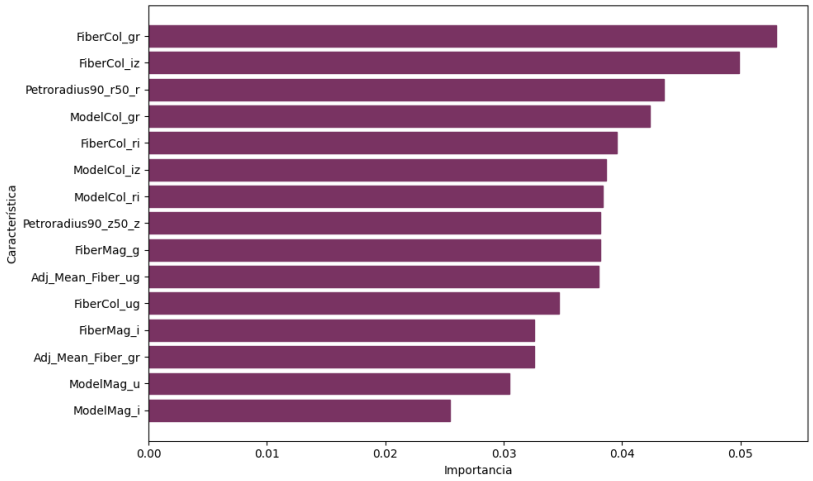

Figura 15: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($1 < z < 2$).

Las Figuras 14 y 15 muestran las características seleccionadas para los intervalos ($0 < z < 1$) y ($1 < z < 2$) respectivamente. Como se puede apreciar, solo en el intervalo ($0 < z < 1$) se pueden distinguir características importantes. Los demás resultados se muestran en el Anexo 6.3.

4.3.1. Eliminación recursiva de características: pruebas

Utilizando las características seleccionadas para cada intervalo, se entrenó el modelo y se realizaron las predicciones de corrimiento al rojo correspondientes. A continuación se presentan los resultados:

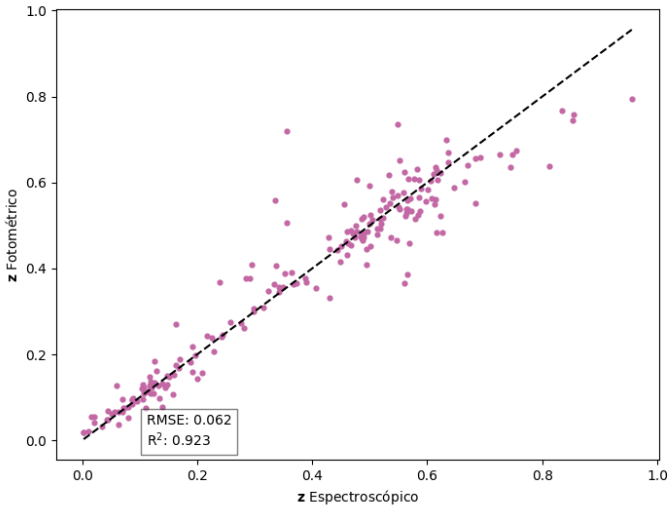

Figura 16: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($0 < z < 1$) utilizando las características seleccionadas por el comando "REF".

La Figura 16 muestra la prueba para el intervalo de ($0 < z < 1$) y se puede apreciar una mejora en las estimaciones, ya que el RMSE disminuyó y el R^2 aumentó, además de que se sigue observando una tendencia hacia la función identidad.

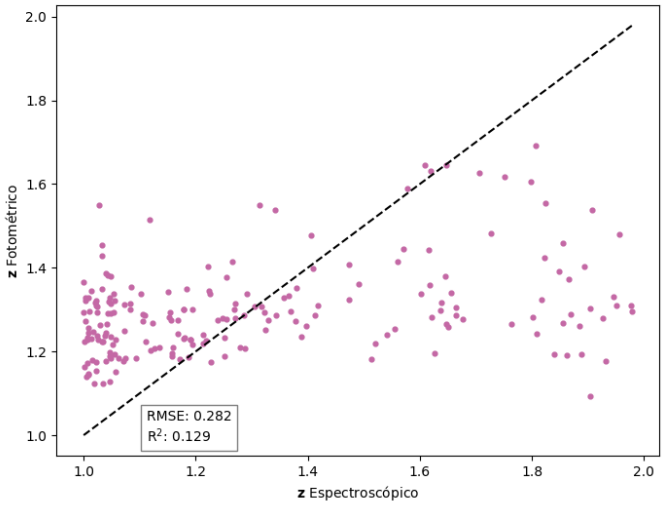

Figura 17: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para para las observaciones en el intervalo ($1 < z < 2$) utilizando las características seleccionadas por el comando "REF".

Sin embargo, en la Figura 17 se muestra como el algoritmo sigue sin poder distinguir o realizar buenas predicciones, ya que el RMSE sigue siendo alto y el valor de R^2 es de 0.033. Esto mismo sucede en los siguientes intervalos mostrados en el Anexo 6.3.1.

4.4. Dominio logarítmico

Ahora se muestran los resultados de aplicar la transformación logarítmica al dominio descrita en la Sección 3.5. En las Figuras 18 y 19 se presentan las gráficas del corrimiento al rojo con la transformación logarítmica contra el corrimiento al rojo en su escala original. Las demás gráficas se encuentran en el Anexo 6.4.

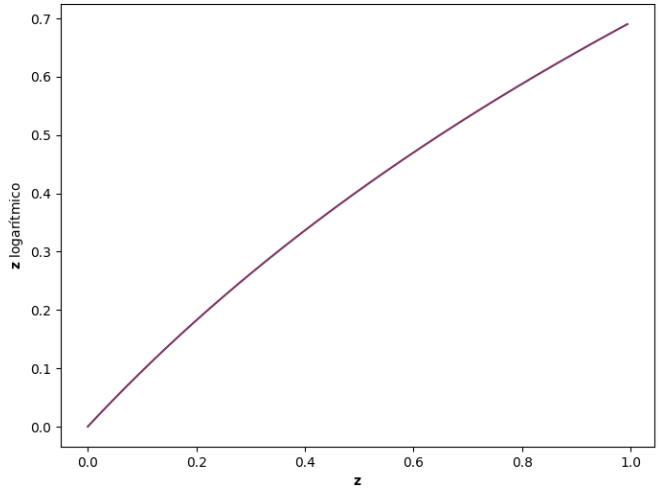

Figura 18: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($0 < z < 1$).

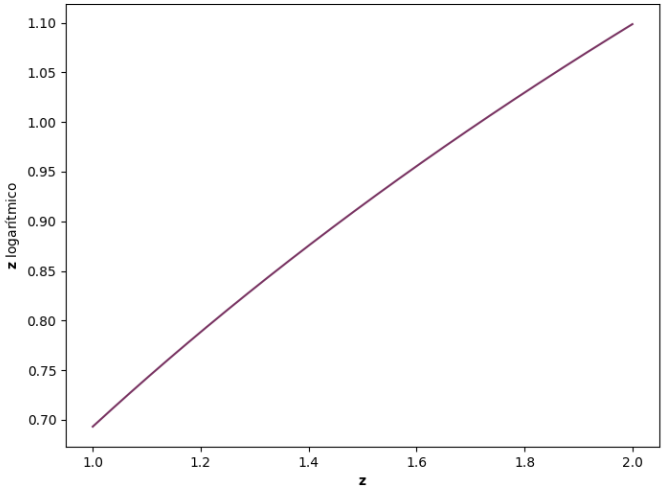

Figura 19: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($1 < z < 2$).

4.4.1. Dominio logarítmico: pruebas

Después de aplicar la transformación logarítmica, se presenta el resultado de realizar predicciones de corrimientos al rojo entrenando al modelo con la escala logarítmica y luego, mediante una transformación exponencial, regresando los datos a su escala original. Las Figuras 20 y 21 muestran los resultados para los intervalos $(0 < z < 1)$ y $(1 < z < 2)$ y en el Anexo 6.4.1 se muestran los resultados para todos los demás.

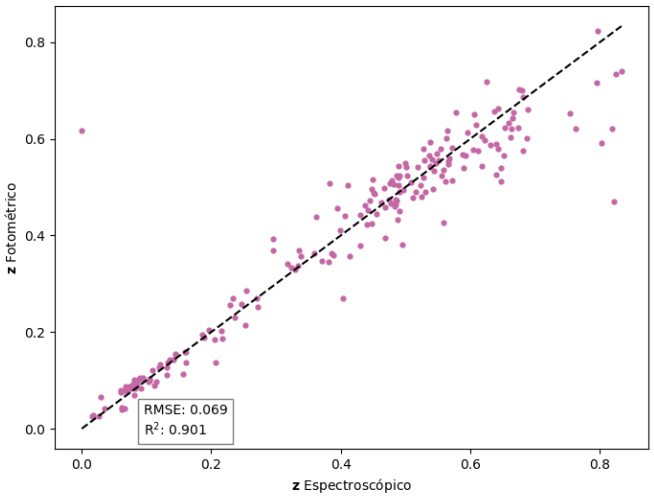

Figura 20: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo $(0 < z < 1)$. Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

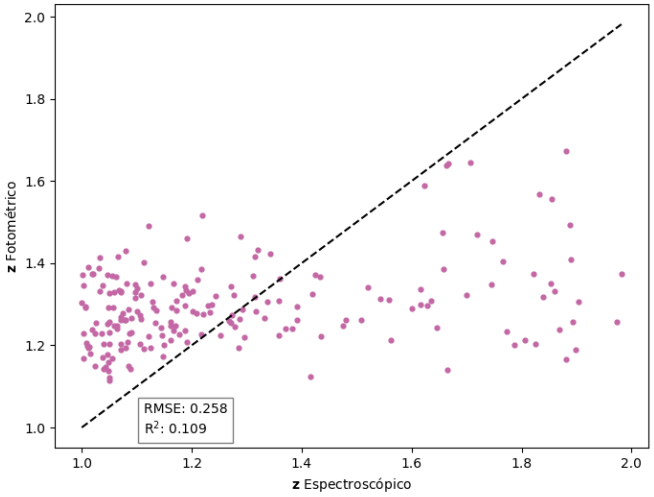

Figura 21: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo $(1 < z < 2)$. Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

De manera similar a las otras pruebas, el algoritmo muestra un buen rendimiento en la estimación de corrimientos al rojo en el intervalo de z: 0 - 1, con un RMSE de 0.062 y un R^2 de 0.923. Sin embargo, para el intervalo de z: 1 - 2, no se observa una tendencia clara, y las métricas de un RMSE de 0.282 y un R^2 de 0.129 indican que el modelo tiene dificultades para realizar predicciones precisas en este rango. Las pruebas en los demás intervalos arrojan un resultado similar a este último.

4.5. Grid Search

En esta sección se exponen los resultados de las predicciones obtenidas al estimar el modelo utilizando los parámetros óptimos encontrados tras aplicar Grid Search.

```
Mejores parámetros encontrados para z: 0-1:
{'max_depth': 15, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 200}
```

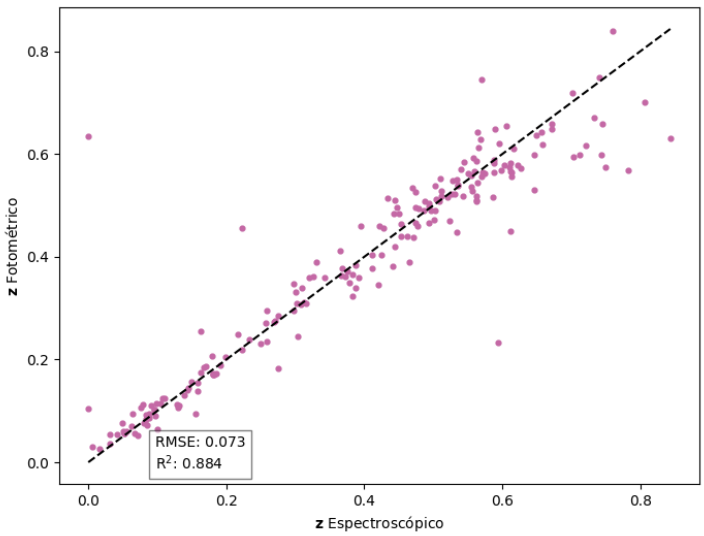

Figura 22: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($0 < z < 1$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

```
Mejores parámetros encontrados para z: 0-1:
{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
```

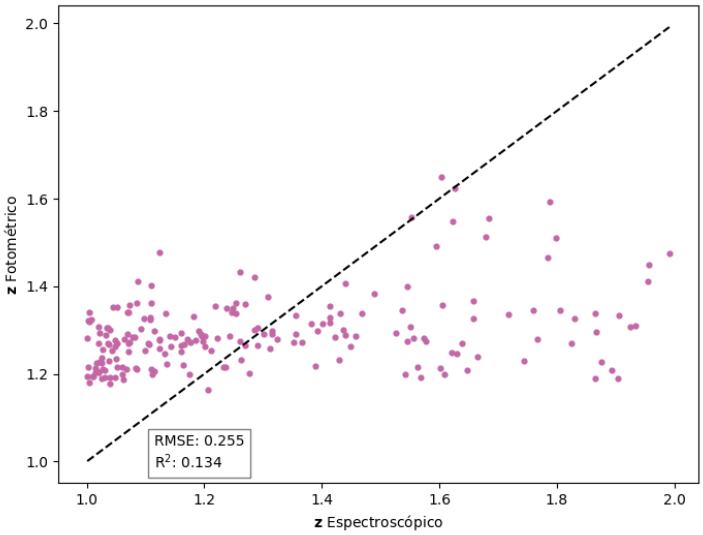

Figura 23: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($1 < z < 2$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

Una vez entrenado el modelo con los parámetros óptimos, para el primer intervalo de corrimiento al rojo se obtuvieron buenas predicciones, ya que con un RMSE de 0.073 y un R^2 de 0.884 se puede concluir que el modelo tiene un rendimiento satisfactorio en este intervalo. En cambio, para el segundo intervalo, sigue sin observarse una tendencia clara, y las métricas de RMSE igual a 0.225 y R^2 de 0.134 indican que el modelo tiene dificultades para hacer predicciones precisas en este rango. Los resultados para los demás intervalos se muestran en el Anexo 6.4.2 y son similares a los del intervalo ($1 < z < 2$).

4.6. Resultado Bootstrap

Los resultados finales se presentan aplicando la técnica de remuestreo Bootstrap, la cual nos permite obtener intervalos de predicción al 95 %. Estos intervalos nos indican en qué rango caerán las predicciones aproximadamente el 95 % de las veces si se repite el proceso de muestreo muchas veces, en este caso, 1000 repeticiones. Además, se presentan las métricas promedio, RMSE y R^2, las cuales nos proporcionan una idea de la precisión y el ajuste del modelo en promedio sobre todas las repeticiones.

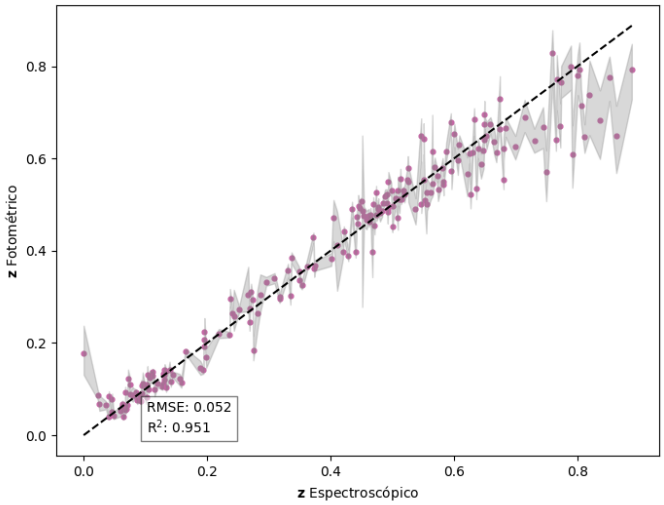

Figura 24: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($0 < z < 1$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

Para el intervalo de z: 0 - 1 y considerando las 1000 repeticiones, los intervalos de confianza sugieren que en la mayoría de las ocasiones las predicciones están cerca de los valores teóricos. Esto se ve respaldado por un RMSE promedio de 0.052 y un R^2 de 0.951. Estos resultados se muestran en la Figura 24.

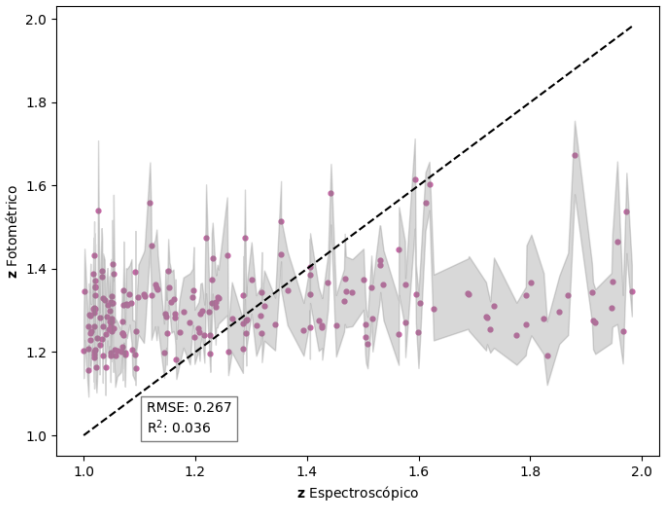

Figura 25: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($1 < z < 2$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

Para el intervalo z: 1 - 2, en la Figura 25 aún se observa una falta de tendencia alrededor de la función identidad. El RMSE = 0.267 y R^2 de 0.036 indican que en la mayoría de las predicciones, el modelo produce resultados que están alejados del valor real. Los demás resultados se presentan en el Anexo 6.4.3.

5. Conclusiones

Este trabajo de tesis consistió en la implementación del algoritmo ExtraTreesRegressor, utilizado en aprendizaje automático para estimar valores numéricos, con el fin de estimar corrimientos al rojo fotométricos de galaxias y cuásares. Los datos se obtuvieron del catálogo del Sloan Digital Sky Survey (SDSS) en todo el rango disponible ($0 < z < 7$). Esta implementación incluyó varias técnicas comunes en algoritmos de aprendizaje automático, como la eliminación recursiva de características, GridSearch, estandarización mediante la función MinMaxScaler y un dominio logarítimico para abordar el desequilibrio en la cantidad de datos por intervalo de corrimiento al rojo. En la Sección 4, al comparar nuestros resultados con los del artículo **"Photometric redshift estimation using ExtraTreesRegressor: Galaxias and quasars from low to very high redshifts"** (Reza y Haque, 2020), observamos diferencias significativas para $z > 1$. Para el conjunto de datos completo, obtuvimos un RMSE = 1.852 y R^2 = 0.216 para los datos con repetición, y RMSE = 2.050 y R^2 = 0.045 sin repetición, mientras que el artículo reporta un MSE = 0.66. Gráficamente, nuestra prueba con datos repetidos se acerca más a la gráfica del artículo, y esta prueba tiene mejores métricas considerando que tenemos un 19 % de datos repetidos entre el conjunto de entrenamiento y prueba. Creemos que esto se debe a que la base de datos utilizada en el artículo contiene datos repetidos, lo que podría haber influido en la obtención de un Error Cuadrático Medio bajo.

Con nuestra base de datos limpia, intentamos obtener estimaciones de corrimiento al rojo. Sin embargo, como se puede ver en las gráficas, las estimaciones solo se ajustaron bien en el intervalo de 0 a 1 y más específicamente de 0 a 0.8. Nuestros resultados coinciden con la bibliografía y trabajos previos mencionados en la introducción, que utilizaron datos del SDSS. Se cree que el intervalo de 0 a 0.8 es donde la información contenida en las densidades de flujo de los filtros **u-g-r-i-z** de los objetos astronómicos es más efectiva para determinar corrimientos al rojo fotométricos. Esto se puede apreciar también con otras técnicas como el ajuste de plantillas o modelos (template fitting). A medida que la galaxia está más lejos, los filtros **u-g-r-i-z** dejan de muestrear la emisión estelar características de las galaxias y ello puede afectar la precisión de las estimaciones de corrimientos al rojo fotométricos.

5.1. Trabajo a futuro.

Consideramos que se podría mejorar la efectividad de este algoritmo para corrimientos mayores a 1 si añadimos magnitudes de filtros con longitudes de onda más largas. Existen experimentos que observan a longitudes de onda más largas, y hay observatorios que se dedican a estas observaciones. Sin embargo, dejamos este trabajo para futuras investigaciones para determinar si existe una contraparte para todas las galaxias o qué porcentaje de galaxias podría beneficiarse, ya que actualmente existen bases de datos más pequeñas que podrían limitar este análisis.

6. Anexo

6.1. Pruebas por intervalo

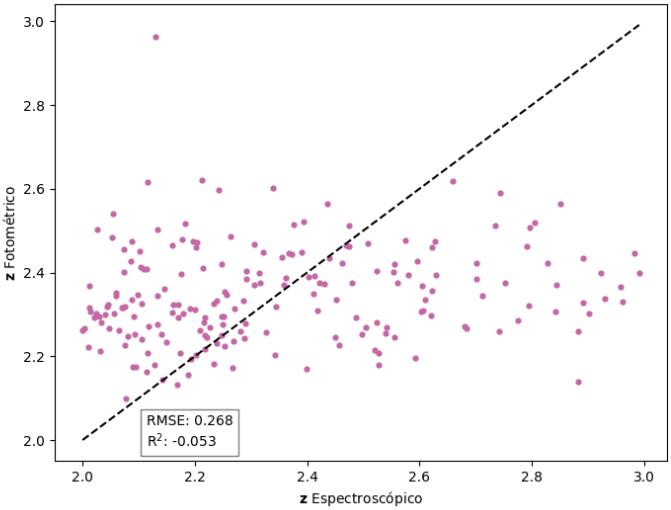

Figura 26: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($2 < z < 3$).

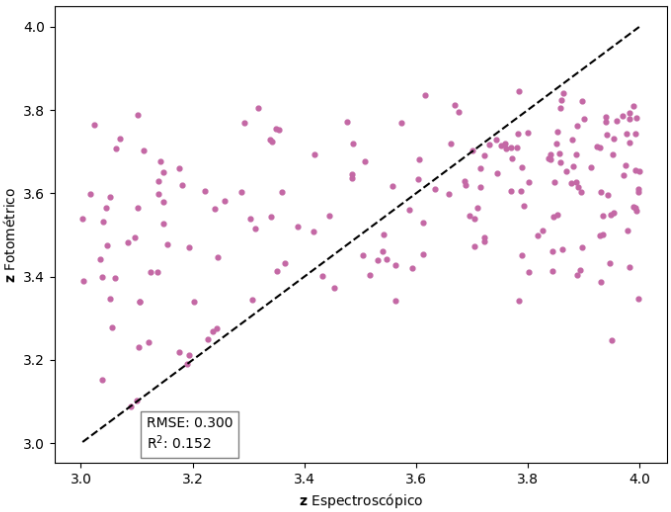

Figura 27: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($3 < z < 4$).

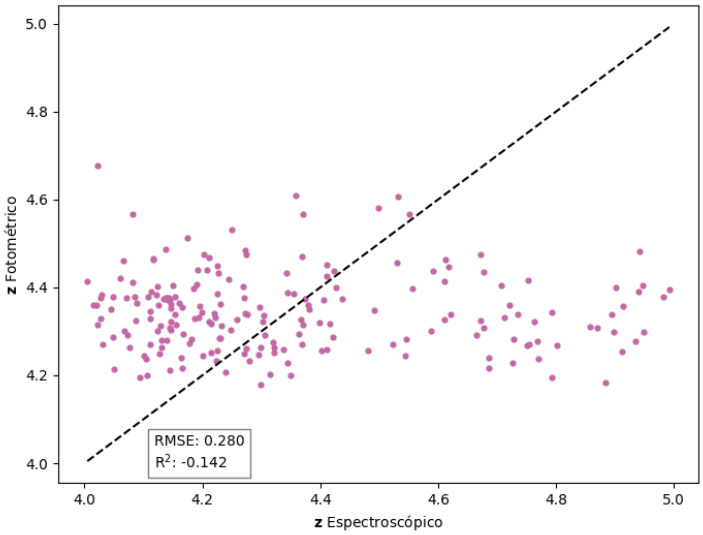

Figura 28: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($4 < z < 5$).

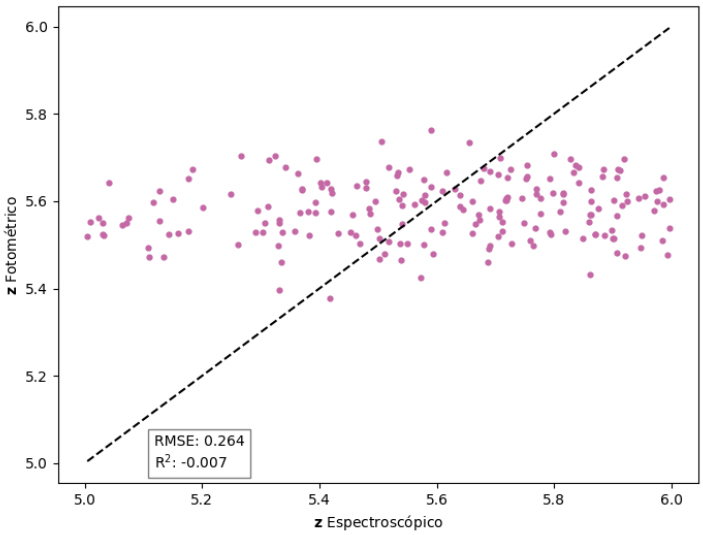

Figura 29: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($5 < z < 6$).

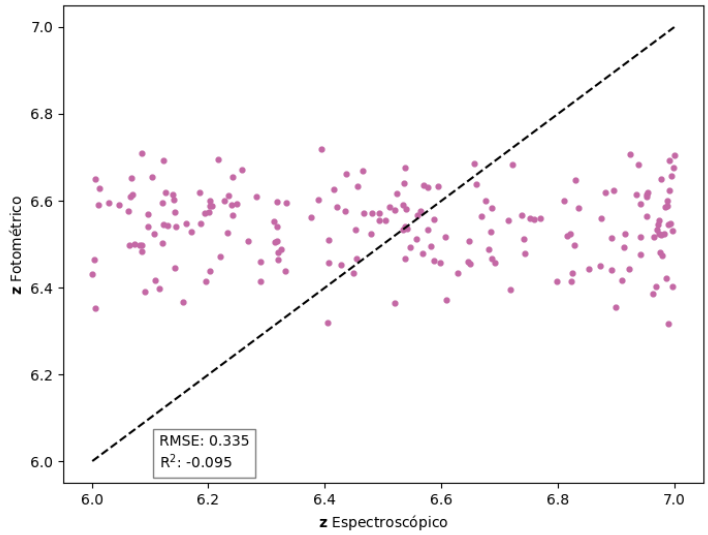

Figura 30: Corrimientos al rojo fotométricos con ExtraTreesRegressor contra corrimientos al rojo espectroscópicos de galaxias y cuásares del SDSS con ($6 < z < 7$).

6.2. Importancia de características

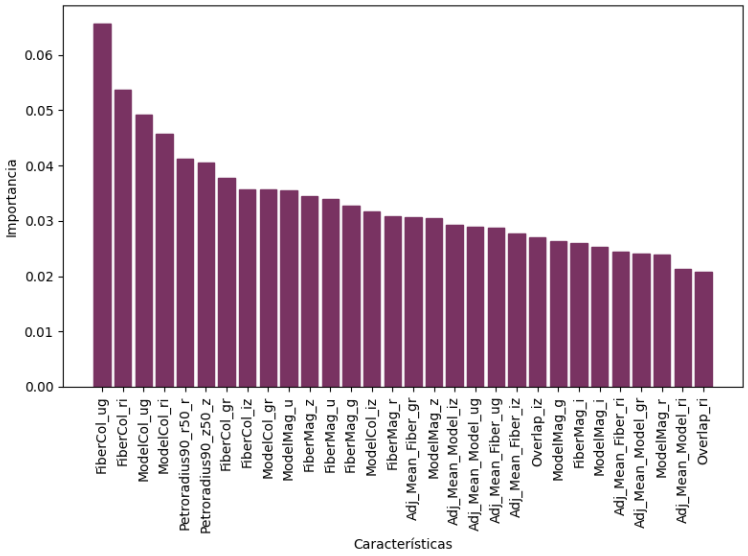

Figura 31: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($2 < z < 3$).

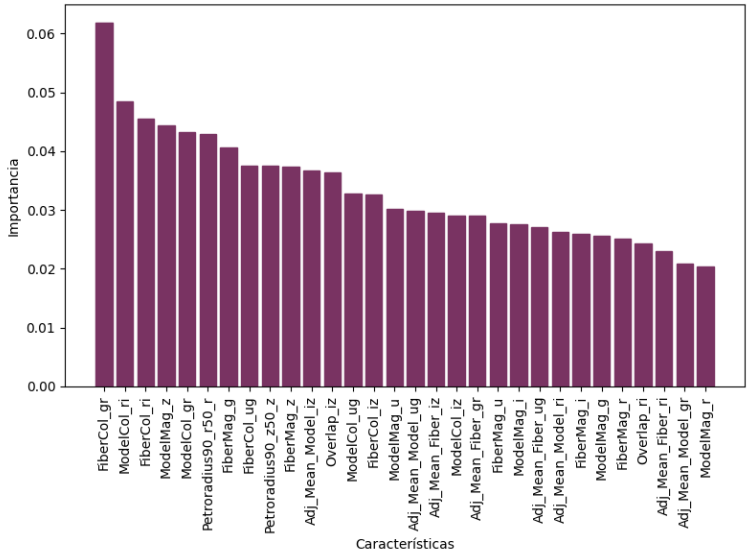

Figura 32: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($3 < z < 4$).

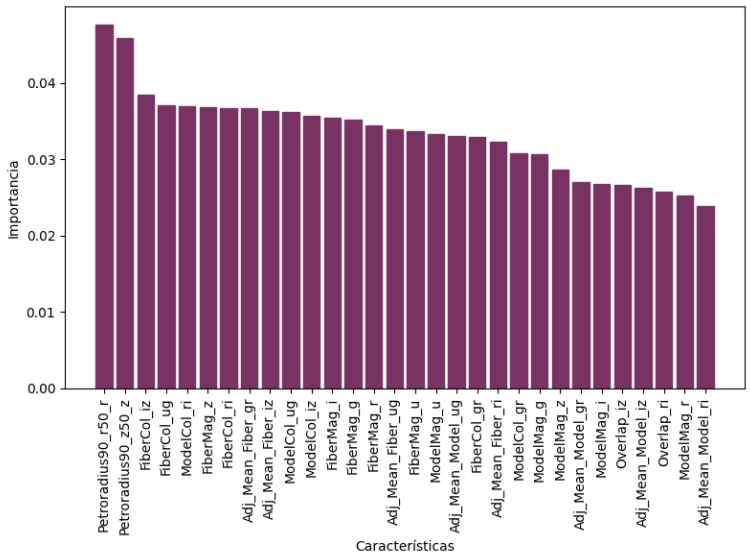

Figura 33: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($4 < z < 5$).

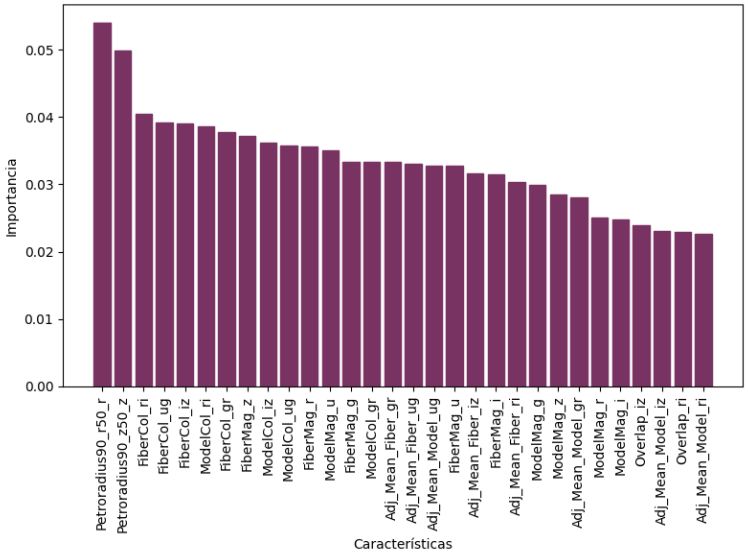

Figura 34: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($5 < z < 6$).

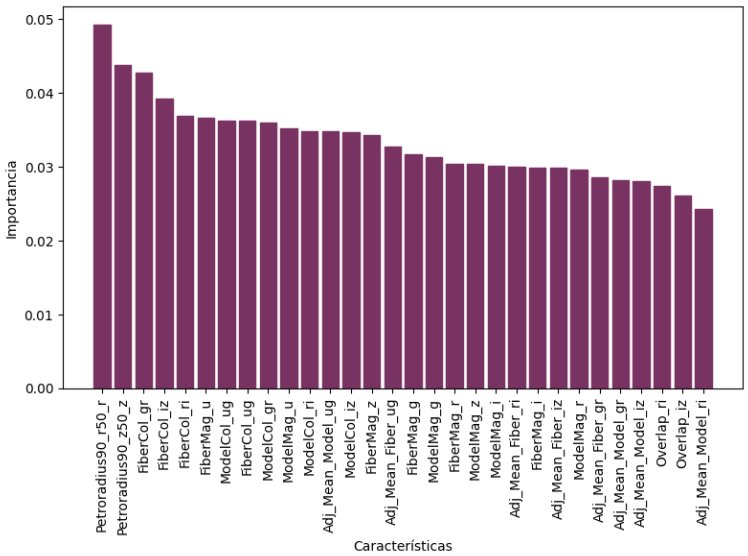

Figura 35: Importancia asignada a cada una de las 30 características utilizadas en este trabajo. La importancia es calculada mediante el comando ‘’feature_importances_” descrito en la Sección 3.3 y aplicado a los datos en el intervalo ($6 < z < 7$).

6.3. Eliminación recursiva de características

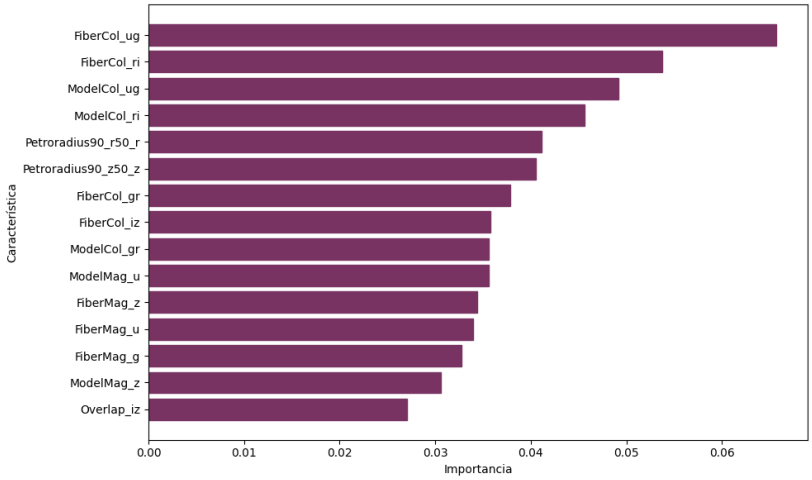

Figura 36: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($2 < z < 3$).

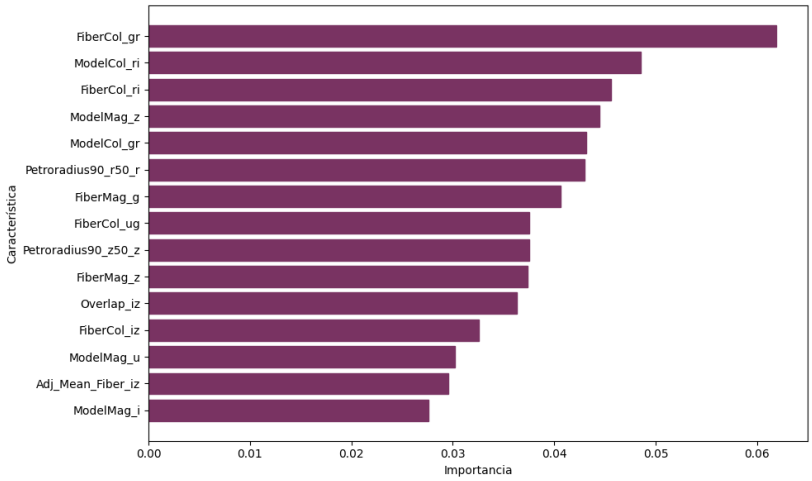

Figura 37: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($3 < z < 4$).

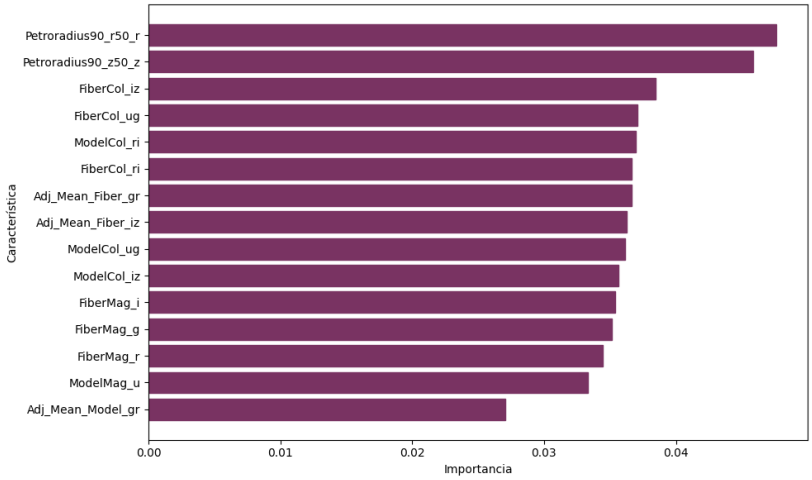

Figura 38: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($4 < z < 5$).

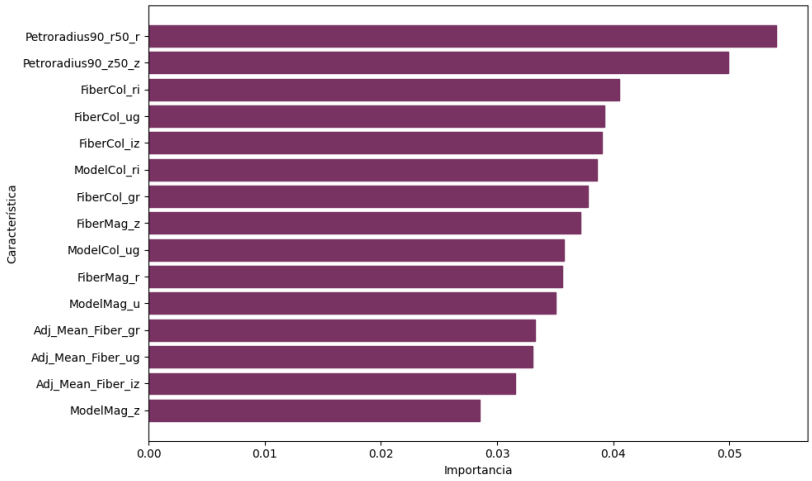

Figura 39: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($5 < z < 6$).

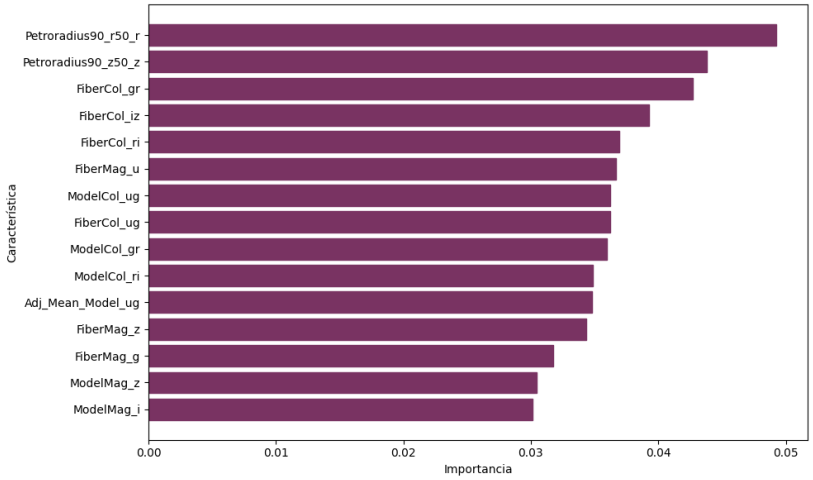

Figura 40: Características seleccionadas mediante el comando ”REF” descrito en la Sección 3.4, aplicado a los datos en el intervalo ($6 < z < 7$).

6.3.1. Eliminación recursiva de características: pruebas

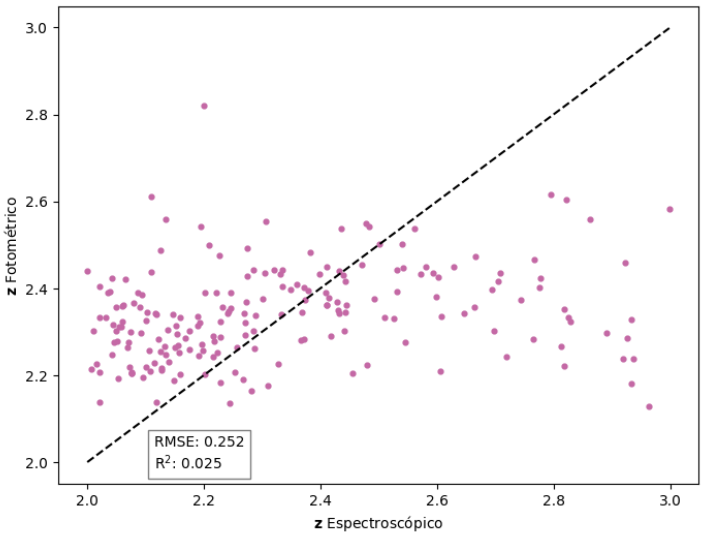

Figura 41: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($2 < z < 3$) utilizando las características seleccionadas por el comando “REF”.

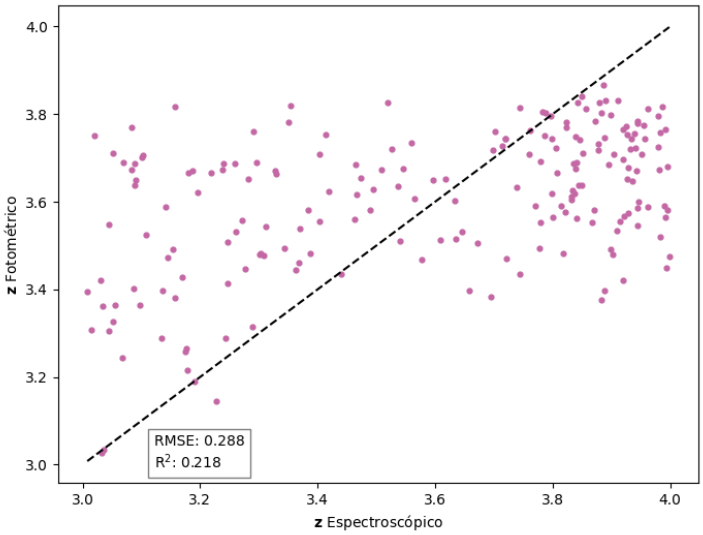

Figura 42: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($3 < z < 4$) utilizando las características seleccionadas por el comando “REF”.

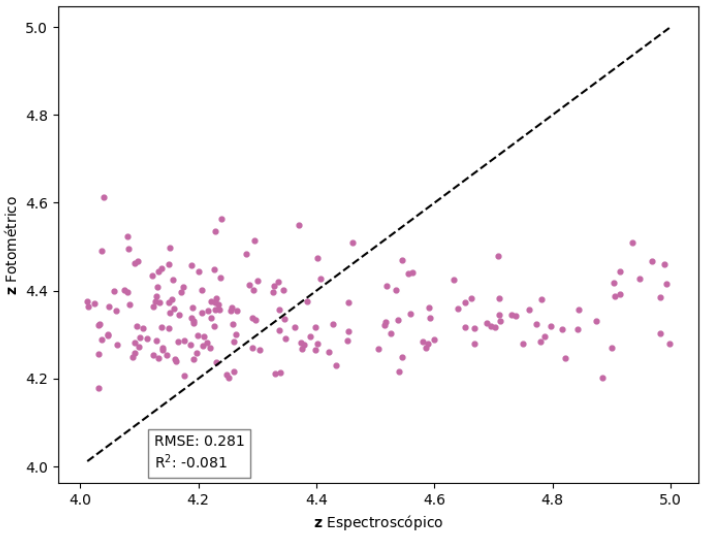

Figura 43: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($4 < z < 5$) utilizando las características seleccionadas por el comando “REF”.

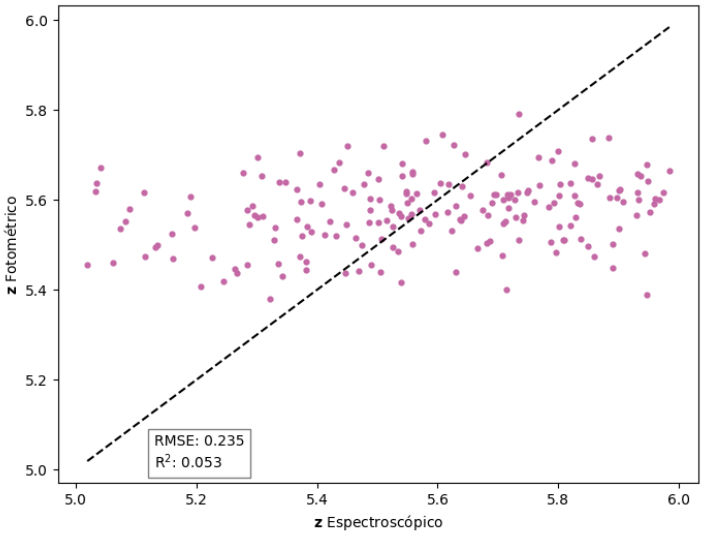

Figura 44: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($5 < z < 6$) utilizando las características seleccionadas por el comando “REF”.

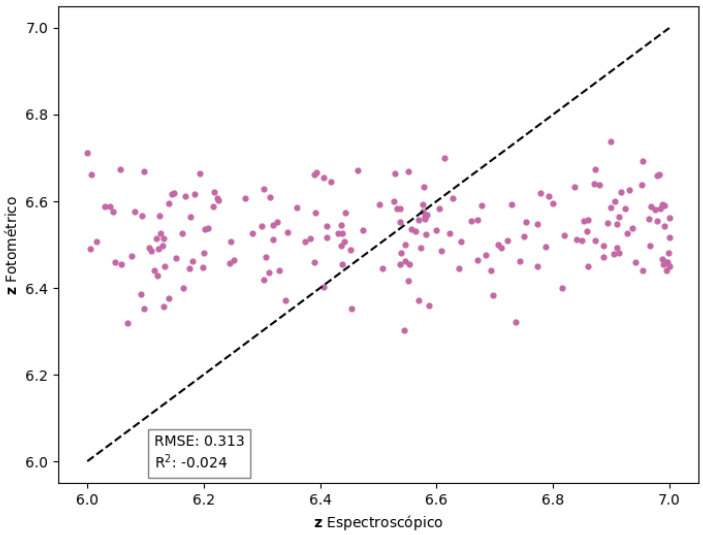

Figura 45: Corrimientos al rojo fotométricos contra corrimientos al rojo espectroscópicos para las observaciones en el intervalo ($6 < z < 7$) utilizando las características seleccionadas por el comando “REF”.

6.4. Dominio logarítmico

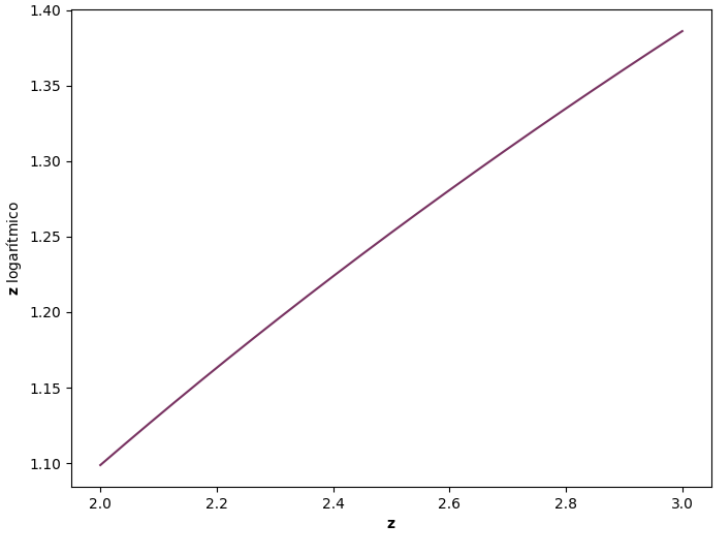

Figura 46: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($2 < z < 3$).

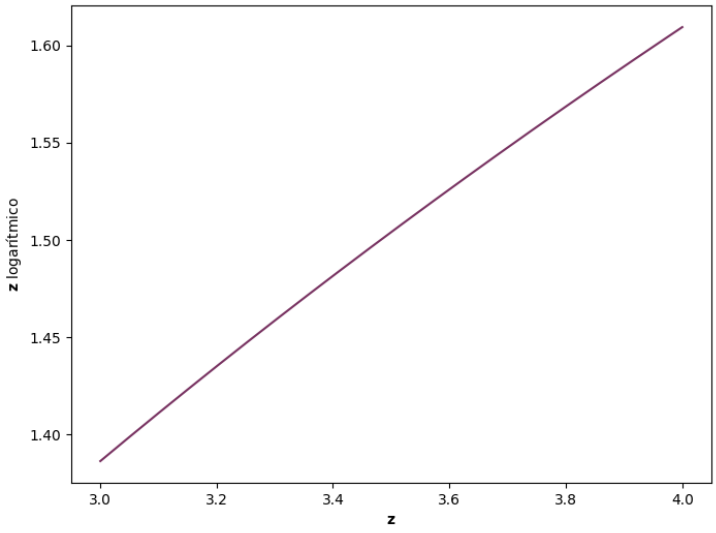

Figura 47: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($3 < z < 4$).

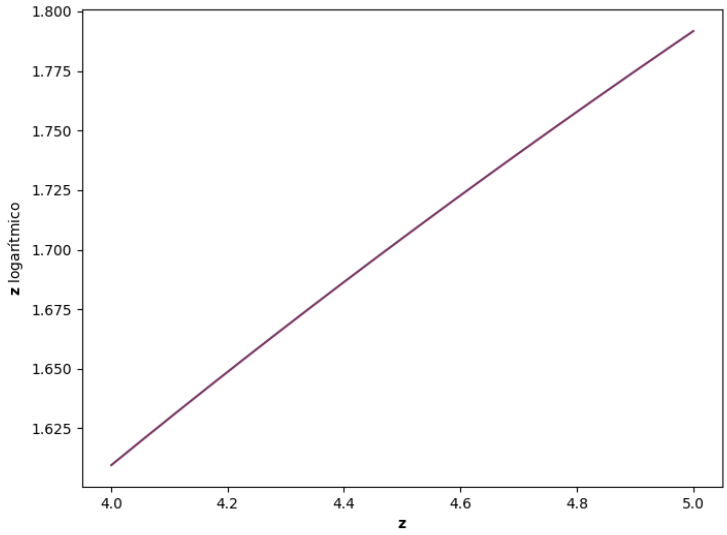

Figura 48: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($4 < z < 5$).

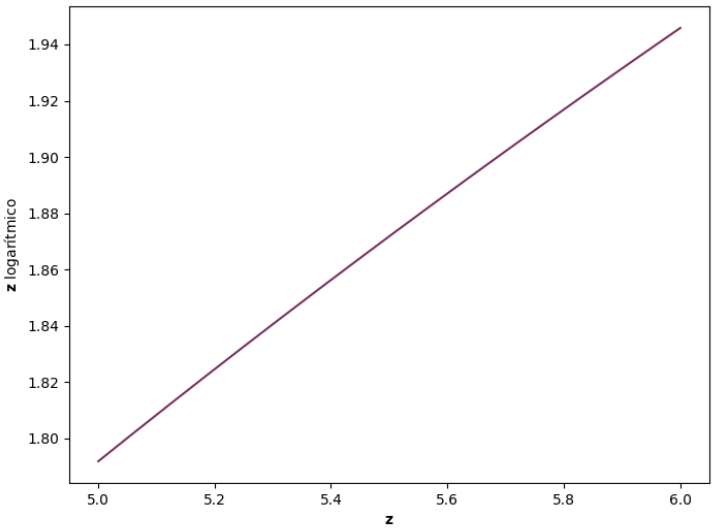

Figura 49: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($5 < z < 6$).

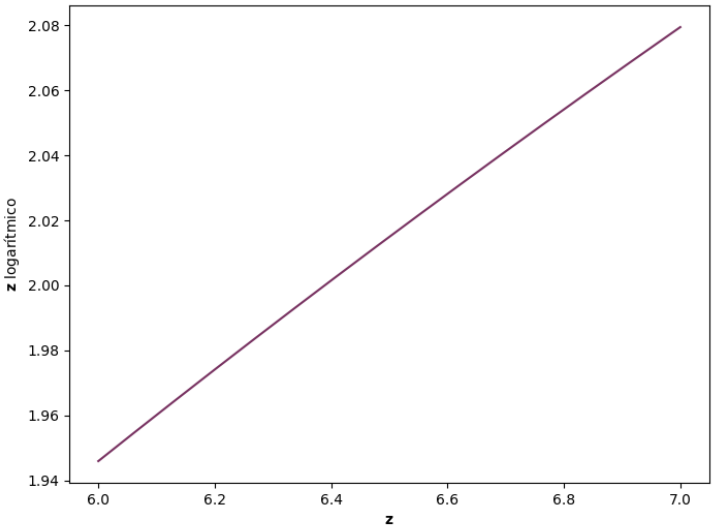

Figura 50: Transformación logarítmica aplicada a los corrimientos al rojo espectroscópicos de la base de datos utilizada para el intervalo ($6 < z < 7$).

6.4.1. Dominio logarítmico: pruebas

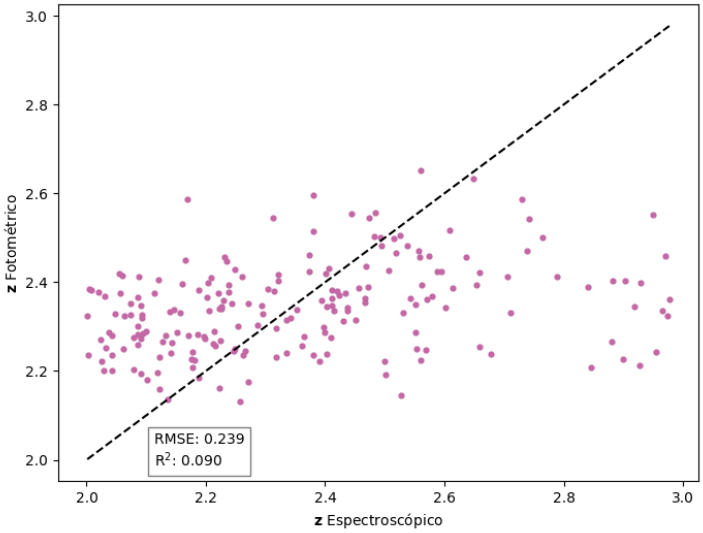

Figura 51: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo ($2 < z < 3$). Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

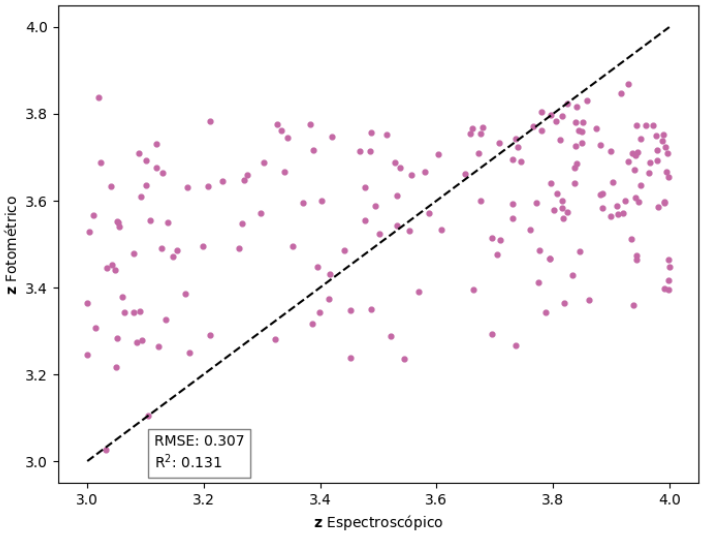

Figura 52: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo ($3 < z < 4$). Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

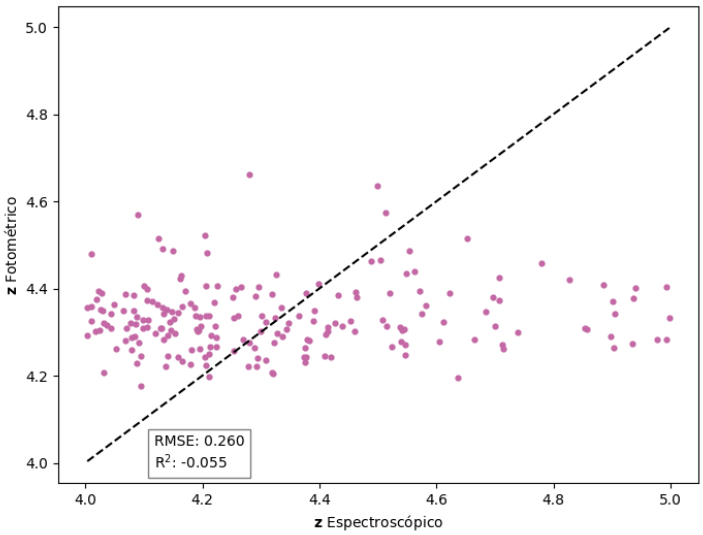

Figura 53: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo ($4 < z < 5$). Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

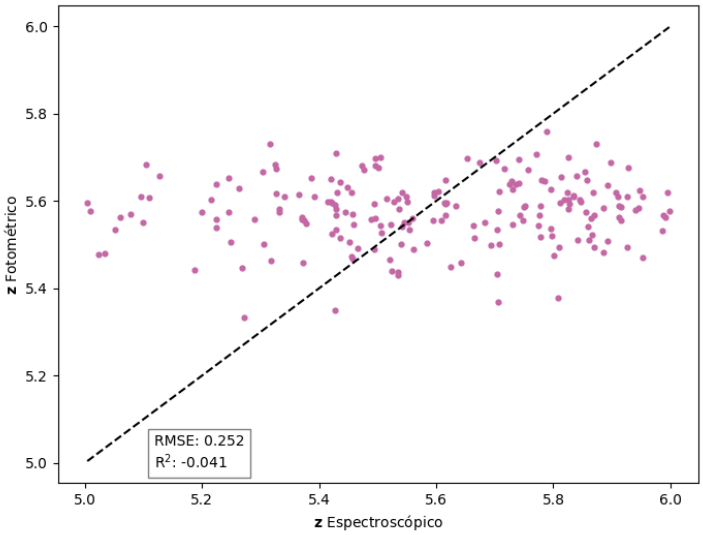

Figura 54: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo ($5 < z < 6$). Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

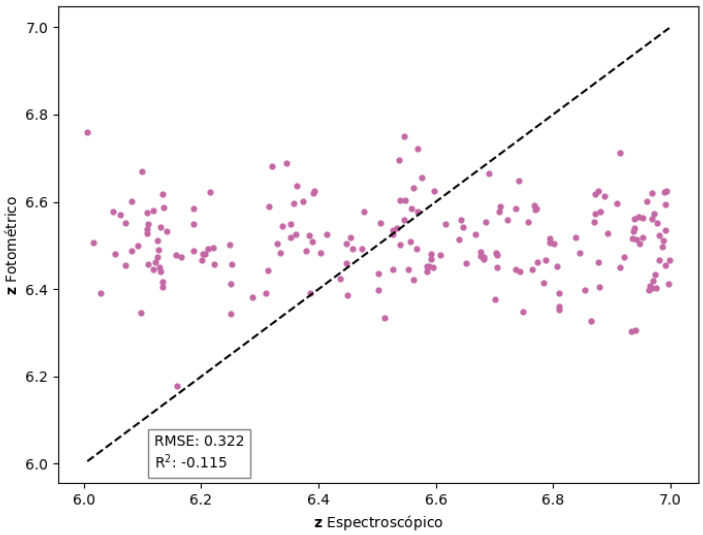

Figura 55: Corrimientos al rojo fotométricos contra espectroscópicos para los objetos en el intervalo ($6 < z < 7$). Los corrimientos al rojo fotométricos se obtuvieron aplicando el dominio logarítmico.

6.4.2. Grid Search

```
Mejores parámetros encontrados para z: 2-3:
{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300}
```

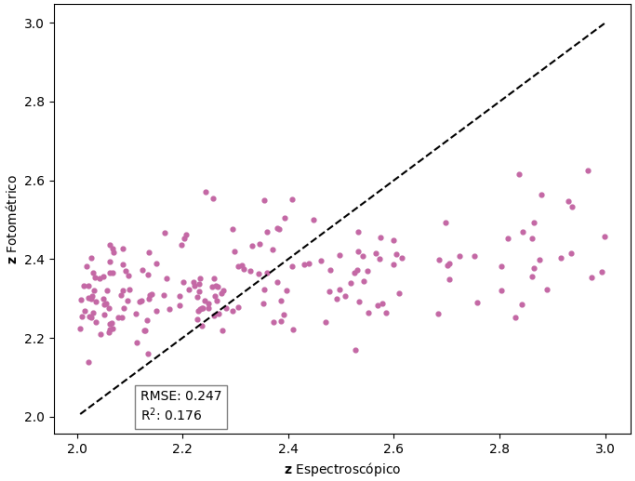

Figura 56: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($2 < z < 3$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

```
Mejores parámetros encontrados para z: 3-4:
{'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 300}
```

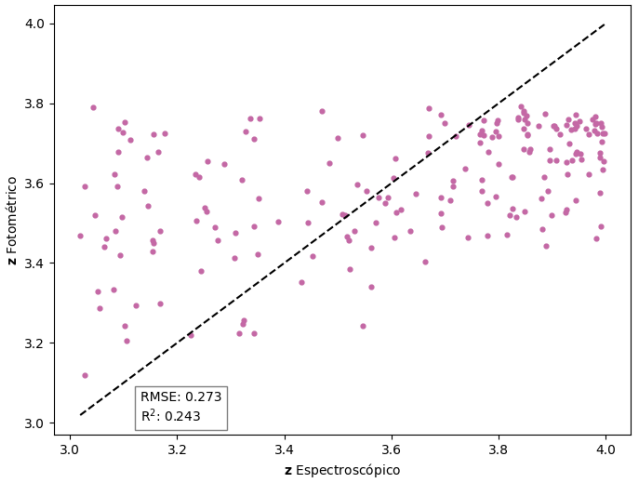

Figura 57: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($3 < z < 4$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

```
Mejores parámetros encontrados para z: 4-5:
{'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 5, 'n_estimators': 300}
```

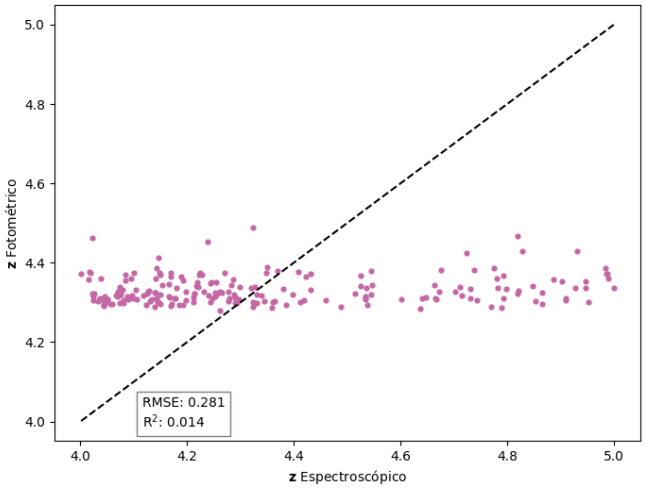

Figura 58: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($4 < z < 5$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

```
Mejores parámetros encontrados para z: 5-6:
{'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200}
```

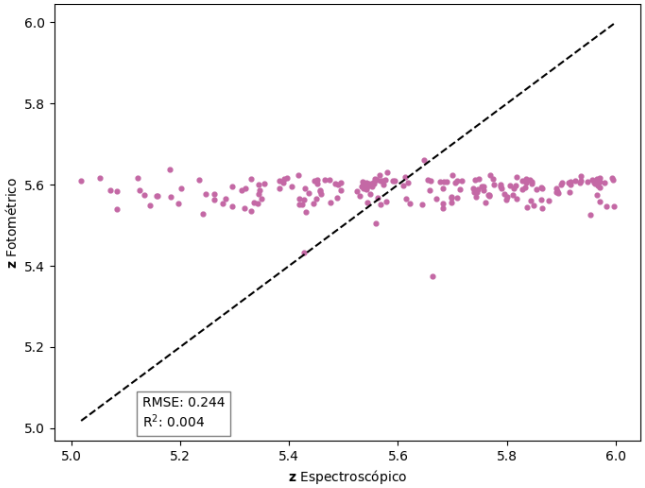

Figura 59: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($5 < z < 6$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

```
Mejores parámetros encontrados para z: 6-7:
{'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100}
```

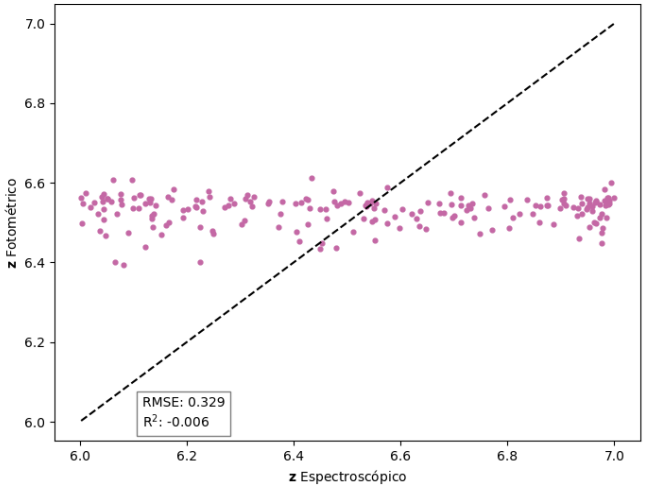

Figura 60: Corrimientos al rojo fotométricos contra espectroscópicos para objetos con ($6 < z < 7$). Los corrimientos al rojo fotométricos se estimaron con los parámetros óptimos seleccionados por el GridSearch.

6.4.3. Resultados Bootstrap

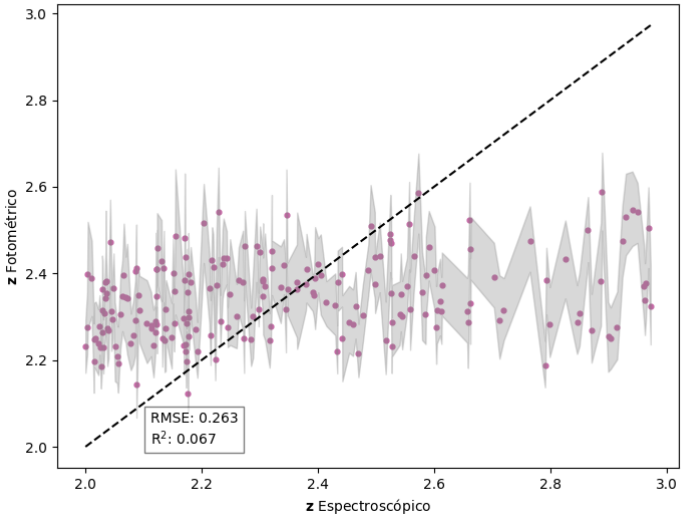

Figura 61: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($2 < z < 3$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

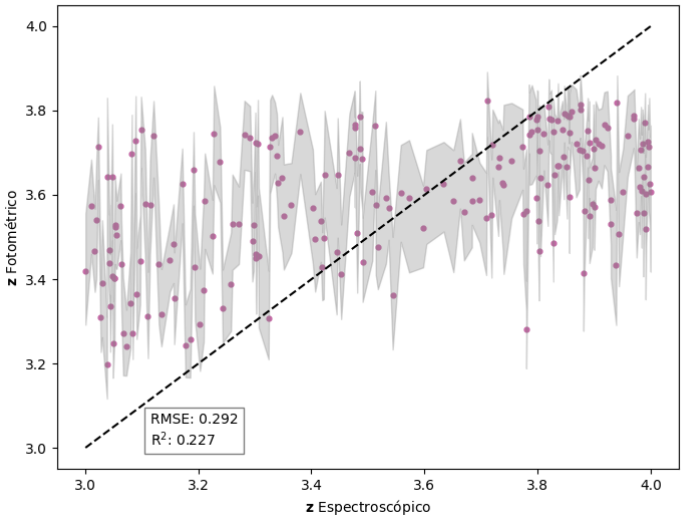

Figura 62: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($3 < z < 4$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

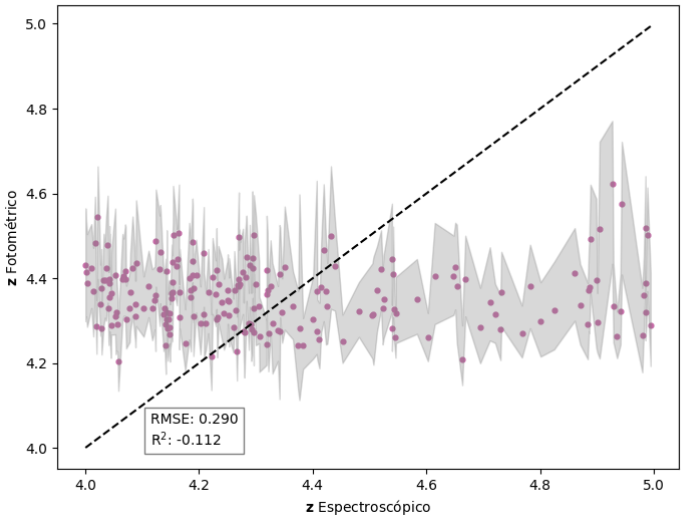

Figura 63: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($4 < z < 5$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

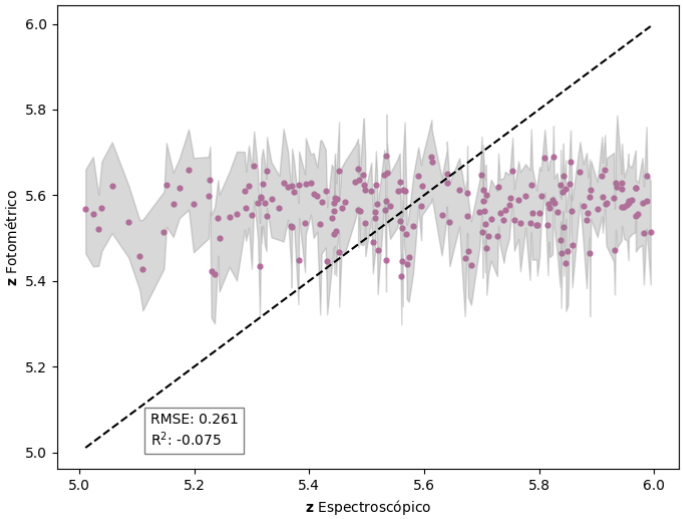

Figura 64: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($5 < z < 6$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

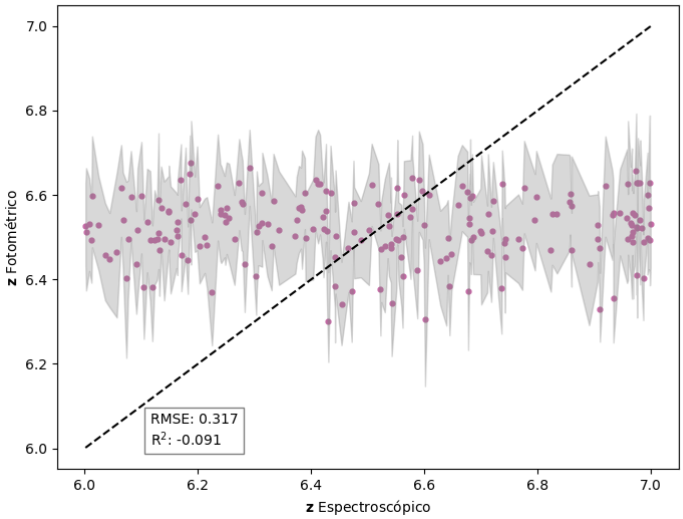

Figura 65: Corrimientos al rojo fotométricos contra espectroscópicos para objetos en el intervalo ($6 < z < 7$). Las estimaciones incluyen bandas de confianza al 95 % obtenidos con la técnica de Bootstrap (área gris).

7. Referencias

Alam, S., Albareti, F. D., Prieto, C. A., Anders, F., Anderson, S. F., Anderton, T., ... Zhu, G. (2015, jul). THE ELEVENTH AND TWELFTH DATA RELEASES OF THE SLOAN DIGITAL SKY SURVEY: FINAL DATA FROM SDSS-III. *The Astrophysical Journal Supplement Series*, *219*(1), 12. Descargado de https://doi.org/10.1088%2F0067-0049%2F219%2F1%2F12

Collister, A. A., y Lahav, O. (2004, apr). ANNz: Estimating photometric redshifts using artificial neural networks. *Publications of the Astronomical Society of the Pacific*, *116*(818), 345–351. Descargado de https://doi.org/10.1086%2F383254

D'Isanto, A., y Polsterer, K. L. (2018). Photometric redshift estimation via deep learning-generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, *609*, A111. Descargado de https://doi.org/10.1051/0004-6361/201731326

Geurts, P., Ernst, D., y Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*. Descargado de https://doi.org/10.1007/s10994-006-6226-1

Gomes, Z., Jarvis, M. J., Almosallam, I. A., y Roberts, S. J. (2017). Improving photometric redshift estimation using gpz: size information, post processing, and improved photometry. *Monthly Notices of the Royal Astronomical Society*, *475*, 331–342. Descargado de https://dx.doi.org/10.1093/mnras/stx3187

Gunn, J. E., Siegmund, W. A., Mannery, E. J., Owen, R. E., Hull, C. L., Leger, R. F., ... others (2006). The 2.5 m telescope of the sloan digital sky survey. *The Astronomical Journal*, *131*(4), 2332–2359. Descargado de https://doi.org/10.1086/500975

Harrison, E. R. (2000). *Cosmology*. Cambridge University Press.

Karttunen, H., Kröger, P., Oja, H., Poutanen, M., y Donner, K. J. (Eds.). (2017). *Fundamental astronomy*. Springer Berlin Heidelberg. Descargado de https://doi.org/10.1007%2F978-3-662-53045-0

Paik, A. D. (2021, may). Doppler effect and its application. *Indian Association For The Cultivation Of Science*. Descargado de https://doi.org/10.13140/RG.2.2.13553.61282

Reza, M., y Haque, M. A. (2020, mar). Photometric redshift estimation using ExtraTreesRegressor: Galaxies and quasars from low to very high redshifts. *Astrophysics and Space Science*, *365*(3). Descargado de https://doi.org/10.1007%2Fs10509-020-03758-w

Robert Resnick, K. S. K. J. W. . S. I., David Halliday. (1992). *Physics, vol. i, 4 th. ed.* Wiley.

Rokach, L., y Maimon, O. (2005). Decision trees. En (Vol. 6, pp. 165–192). Descargado de https://dx.doi.org/10.1007/0-387-25465-X_9

Smith, J. A., Tucker, D. L., Kent, S., Richmond, M. W., Fukugita, M., Ichikawa, T., ... York, D. G. (2002). The ugriz standard-star system. *The Astronomical Journal*, *123*(2), 855–860. Descargado de https://doi.org/10.1086/339311

Wadadekar, Y. (2005, jan). Estimating photometric redshifts using support vector machines. *Publications of the Astronomical Society of the Pacific*, *117*(827), 79–85. Descargado de https://doi.org/10.1086%2F427710