Ciencia de Datos

Carbajal Quintana Carolina

Andrés Ledesma Ramírez

2023-2

Pedro Arturo Flores Silva
Karen Rubi Jiménez López

Proyecto final: Análisis y corroboración de los tipos de estrellas y su lugar en el diagrama H-R mediante la aplicación de la regresión lineal para futuros datos

Índice

1. Resumen

2. Objetivo general

Analizaremos los datos de más de 200 estrellas con el propósito de corroborar el lugar que ocupan en el diagrama H-R, y de esta forma, ayudar a la clasificación de futuras estrellas de las cuales se conozcan características como la luminosidad, radio, temperatura, etc.; todo esto con ayuda de la regresión lineal. Por otra parte, demostraremos la dependencia que hay entre varias de estas características solo conociendo sus valores y sin necesidad de comparar la información con la literatura.

3. Hipótesis

El análisis de las estrellas en nuestro dataset nos mostrará que el lugar que ocuparían en el diagrama H-R es el correcto tomando en cuenta las características de cada una de ellas.

4. Introducción

4.1. Estrellas

Una estrella es una esfera muy caliente de gas (mayormente hidrógeno y helio, aunque están compuesta de muchísimos más elementos) unida por su propia gravedad.

Las estrellas tiene colores muy variados pues estos dependen de su edad, composición, distancia y temperatura. Estrellas frías tienen colores más rojizos y las más calientes tiene colores azules.

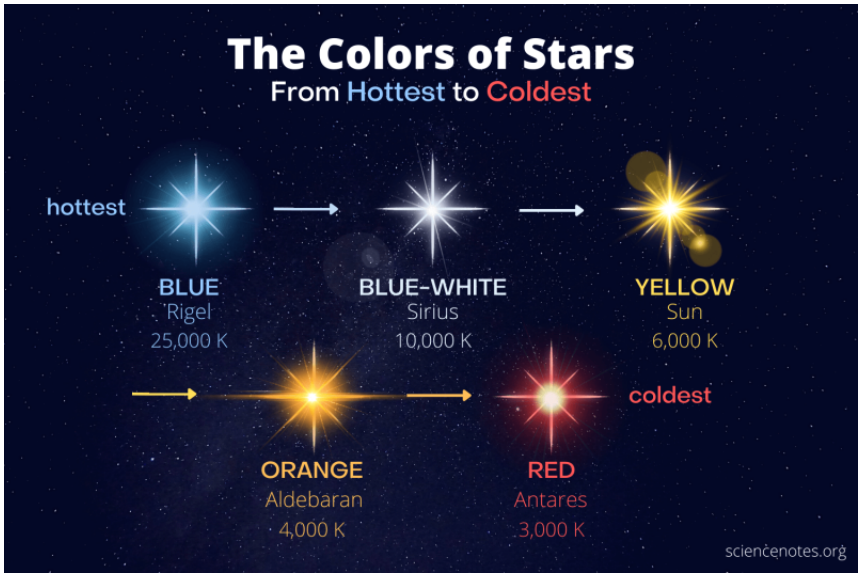

Figura 1: Colores de las estrellas

Por otra parte, las estrellas pueden ser clasificadas por sus temperaturas de superficie, determinadas a partir de la ley de desplazamiento de Wienn (que establece una relación entre la longitud de onda en la que se produce el pico de emisión de un cuerpo negro y la temperatura).

Para una determinada gama de temperaturas sólo se pueden observar determinadas líneas de absorción, ya que sólo en ese rango son poblados los niveles de energía atómica involucrados.

Clase espectral	Temperatura (K)
O	41,000
B	31,000
A	9,240
F	7,240
G	5,920
K	5,300
M	3,850

4.2. Dataset

El dataset que utilizaremos será el de "Star dataset to predict star types"que se puede descargar desde Datos
En este dataset encontramos los valores de temperatura, luminosidad relativa, radio relativo, magnitud absoluta, color, clase y tipo de estrella.

La luminosidad relativa la encontramos calculando $\frac{L}{L_0}$, donde L_o es la luminosidad promedio del Sol (aproximadamente $3{,}828x10^26$ Watts) y L es la luminosidad promedio de la estrella elegida.

El radio relativo se calcula mediante $\frac{R}{R_0}$, donde R es el radio de la estrella elegida y R_0 es el radio promedio del Sol (aproximadamente $6{,}9551x10^8$ m).

De esta forma, el dataset utiliza la ley de Stefan-Boltzmann para la radiación de cuerpo negro, la Ley de desplazamiento de Wienn, el cálculo del radio de una estrella mediante el paralaje y las relaciones de luminosidad y radios relativos.

4.3. Diagrama H-R (HRD)

El diagrama Hertzsprung-Russello diagrama H-R es un gráfico que nos posiciona a las estrellas tomando en cuenta sus magnitudes absolutas o luminosidades en comparación con las temperaturas o las clasificaciones espectrales. Fue creado en 1905 por el astrónomo Ejnar Hertzsprung y, de manera independiente, en 1913 por Henry Norris Russell.

El HRD representa la magnitud absoluta de una estrella en luz visible frente a su tipo espectral, que es una manera de estimar su temperatura. Se hace algo como colocar todas las estrellas a la misma distancia, y representar su brillo frente a su temperatura.

Por convención, el eje horizontal del diagrama recorre las temperaturas de mayor a menor, mientras que el eje vertical recorre los brillos de menos brillante a más brillante.

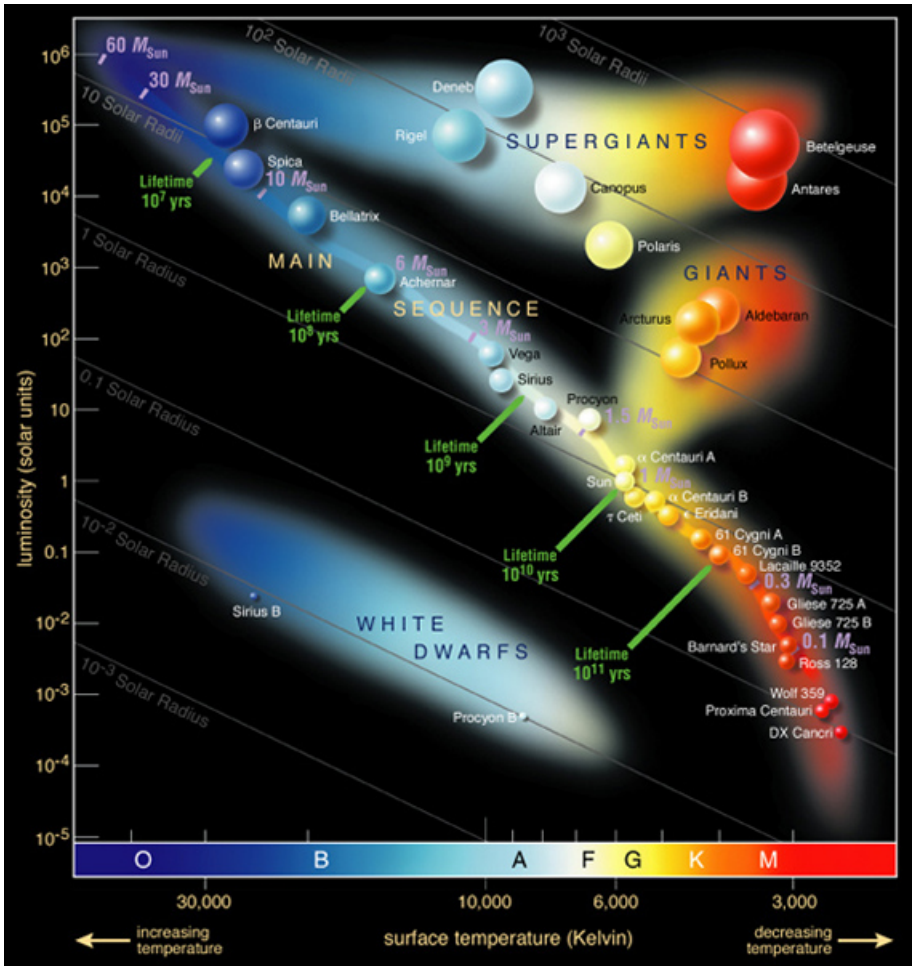

Figura 2: Diagrama H-R

4.4. Regresión lineal

La regresión lineal es un modelo matemático que se utiliza para predecir el valor de una variable respecto a otra.
La regresión lineal se ajusta a una línea recta o a una superficie que minimiza los errores entre los valores de salida previstos y reales.
En palabras más sencillas, consiste en predecir un parámetro (Y) a partir de un parámetro conocido X.

5. Resultados (análisis y discusión)

5.1. Plots informativos

Como primer paso, elaboramos ciertas gráficas para conocer un poco más acerca de los datos con los que estaremos trabajando.
En el primer gráfico (que es un diagrama de dispersión) 3 podemos ver que hay cierta relación entre la magnitud absoluta de una estrella y la temperatura que ésta posee.
El color nos refleja el tipo de estrella mientras que el tamaño es el reflejo de los radios relativos.

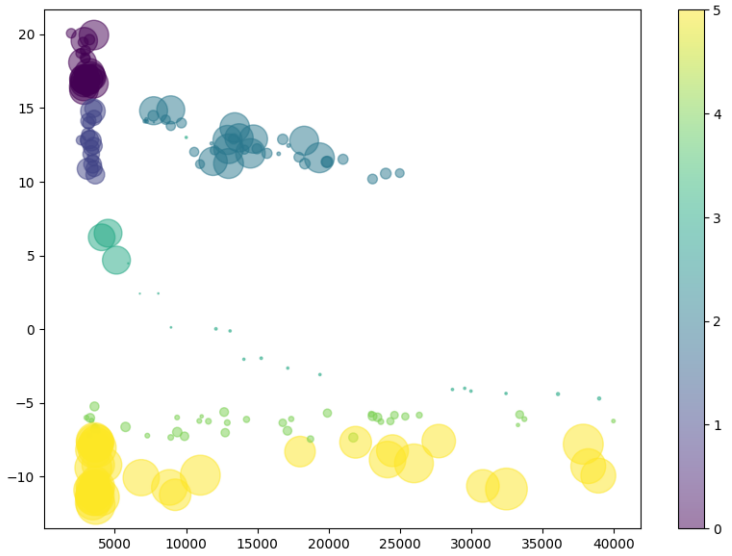

Figura 3: Gráfico de magnitud absoluta vs temperatura

Realizamos además un heatmap que nos muestra las relaciones que hay entre todas las variables con las que estamos trabajando.

Encontrar un 1.0 (recuadro amarillo), nos indica que es el punto de encuentro con la misma variable en ambos ejes, y que el número vaya cambiando conforme las variables van pasando, nos indica la "fuerza" en la relación que tiene esas variables (como la poca fuerza entre la magnitud absoluta y la temperatura, pero la alta relación entre la luminosidad y el tipo de estrella).

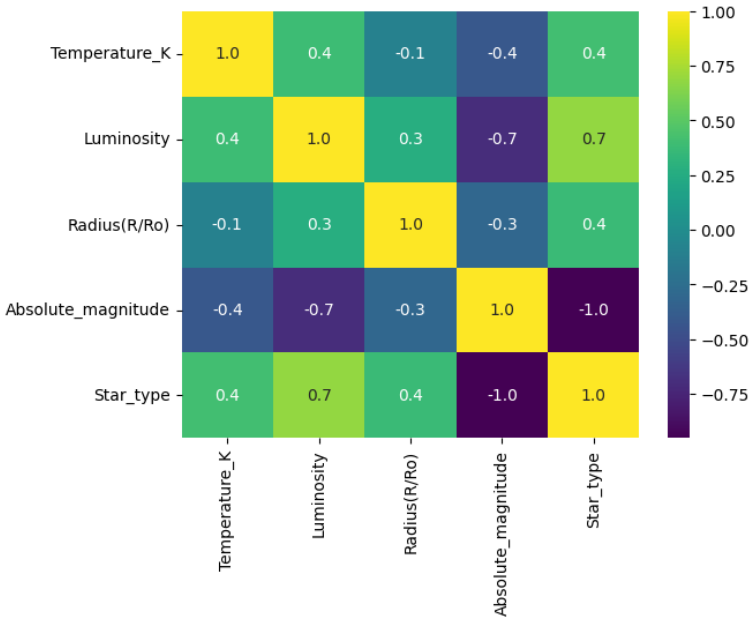

Figura 4: Heatmap de las variables en el dataset

Por otra parte, en la figura 5, vemos 5 gráficas con informaciones muy importantes.
Podemos resaltar aquella que nos muestra que hay 40 estrellas por cada tipo (tipo 0, tipo 1,...,tipo 5) y que se encuentra separada en el gráfico 6

Además, tenemos el número de estrellas en cada magnitud absoluta, así como la temperatura que predomina (en promedio) en todo nuestro dataset (primera gráfica de izquierda a derecha).

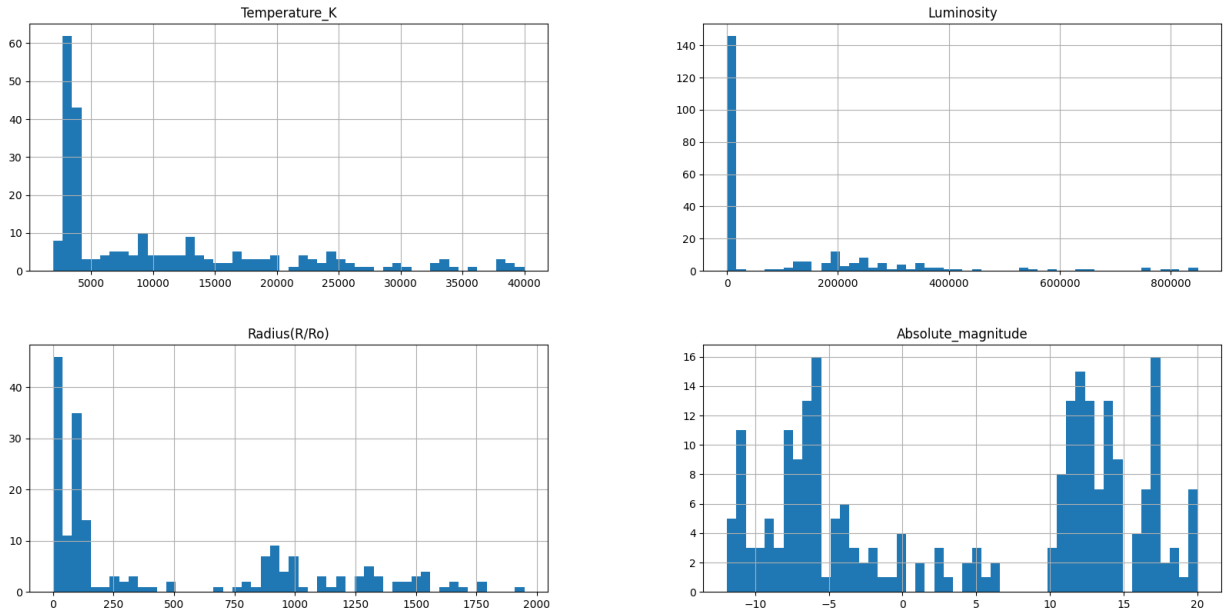

Figura 5: Gráfico de magnitud absoluta vs temperatura

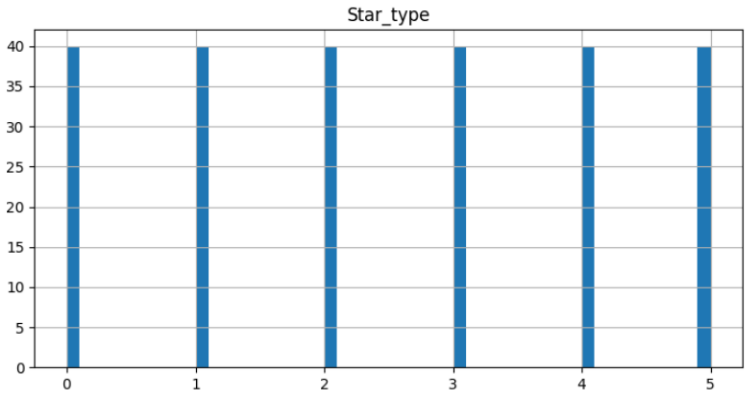

Figura 6: Gráfico la cantidad de estrellas por tipo

Gracias a las gráficas anteriores, notamos entonces que abundan las estrellas con temperaturas menores a los 5000K y poca luminosidad (a partir del tipo G hasta el tipo M), tal y como nos lo corrobora el diagrama H-R que ya conocemos.
Además, hasta el momento, vemos que hay varias variables que "van de la mano", como la temperatura con la luminosidad y el tipo de estrella, y hay otras que no se llevan tan bien (como la magnitud absoluta con el resto, esto pues la magnitud relativa es una variable que no depende de qué tan brillante o caliente es la estrella).

5.2. Entrenamiento y prueba

Para el análisis con regresión lineal, elegimos un 80 % para train y un 20 % para test.

Después de pasar por un proceso de cálculos de pendientes, transformaciones de ciertas variables y de verificar que fuesen los datos que requeríamos, obtuvimos las siguientes gráficas en las cuales observamos la relación

entre las variables con las que se entrenó nuestro modelo y las variables prueba que vendrían siendo el comportamiento del modelo si se le aplicaran nuevos datos.

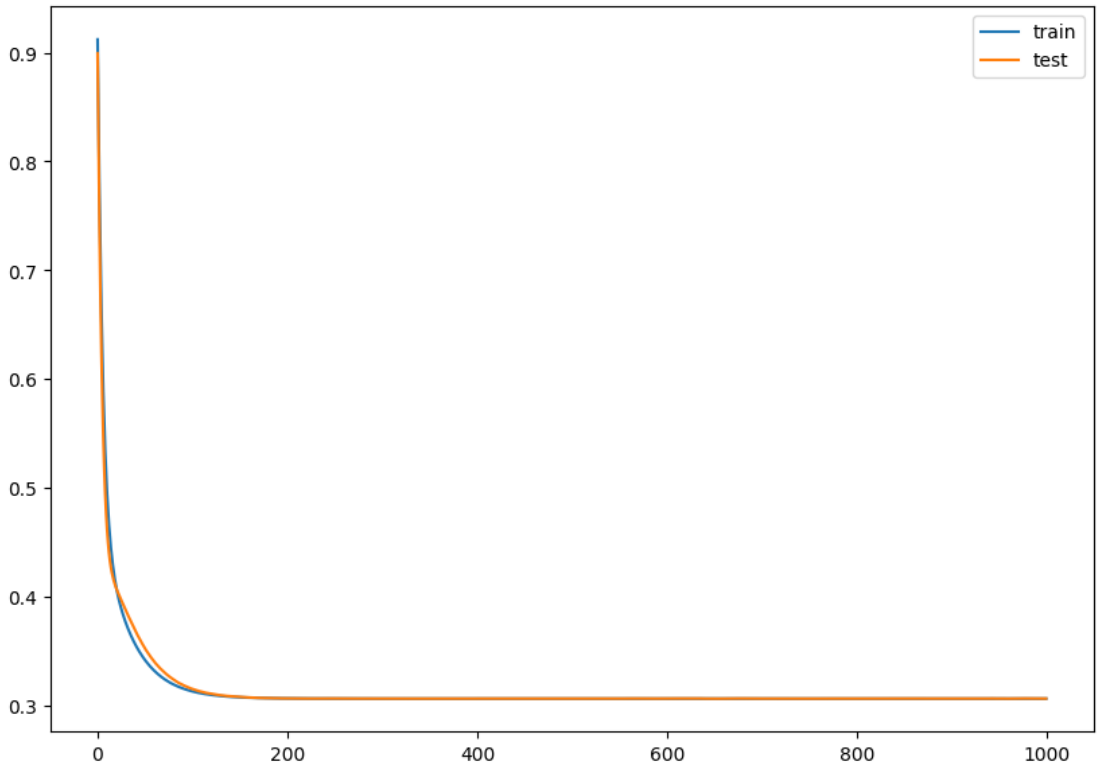

Figura 7: Primer gráfico de supuestos

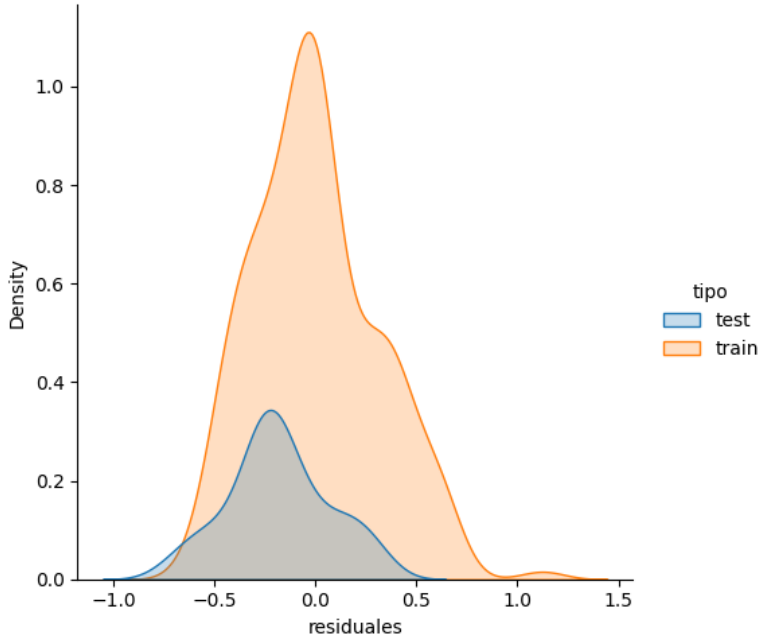

Figura 8: Segundo gráfico de supuestos

En los gráficos 7 y 8 podemos ver que los datos de prueba siguen una forma muy parecida a la de los datos de entrenamiento, y esto es muy buena señal, pues nos indica que el comportamiento del modelo seguiría siendo correcto si se da el caso en el que necesitemos aplicarle aún más variables de las que ya tenemos.

Las líneas con color naranja nos muestran ese 80 % con el que se entrenó y las líneas de color azul nos muestran ese 20 % con el que estamos probando si todo funciona correctamente.

Por otro lado, tenemos un par de gráficos de probabilidad en los que notamos que las variables de test y train se acercan a la pendiente en color rojizo. 9 10
Ambas gráficas nos muestran que hay un poco de dispersión respecto a esa recta; sin embargo, es muy parecida en ambos casos, por lo que sería otra forma de verificar el buen entrenamiento que se le dió al modelo (ya que una gráfica es de las variables que lo entrenaron y otra es de las que lo están probando).

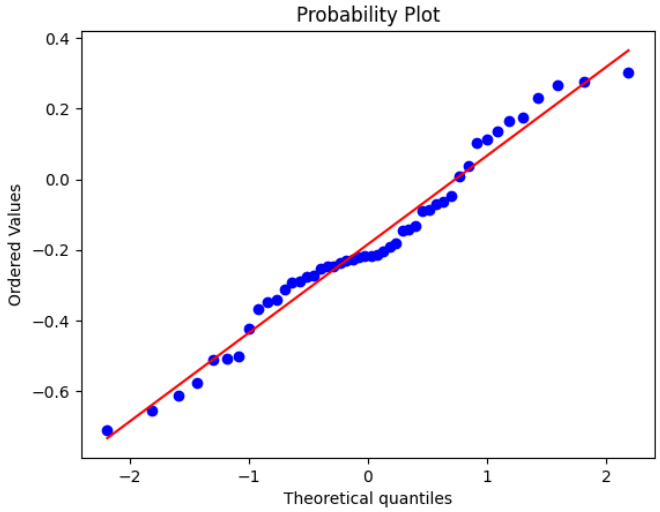

Figura 9: Gráfico de probabilidad para las variables test

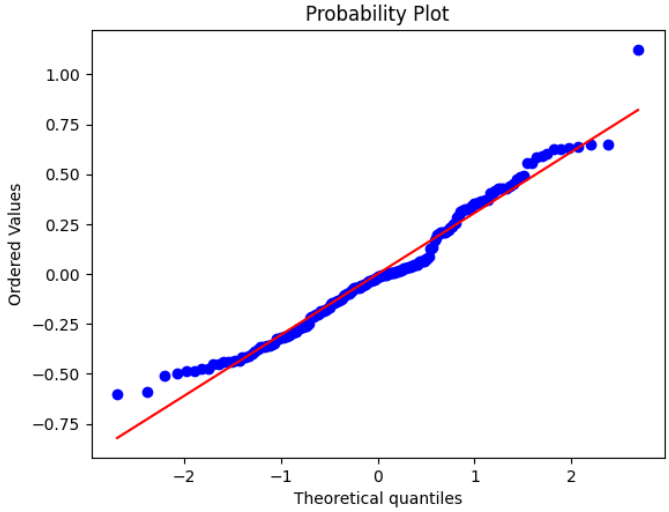

Figura 10: Gráfico de probabilidad para las variables train

Asimismo, nos encontramos con una gráfica de Homocedasticidad 11.

Este gráfico nos muestra que la varianza de los errores es constante a lo largo del tiempo pues en el eje X se encuentran los valores que se predicen y en el eje Y se encuentran sus respectivos “residuos”.
En color naranja tenemos los valores de entrenamiento y en color azul aquellos que se utilizaron para probar el modelo.

En ambos casos se puede ver claramente un comportamiento muy parecido y hasta cierto punto constante en el tiempo.

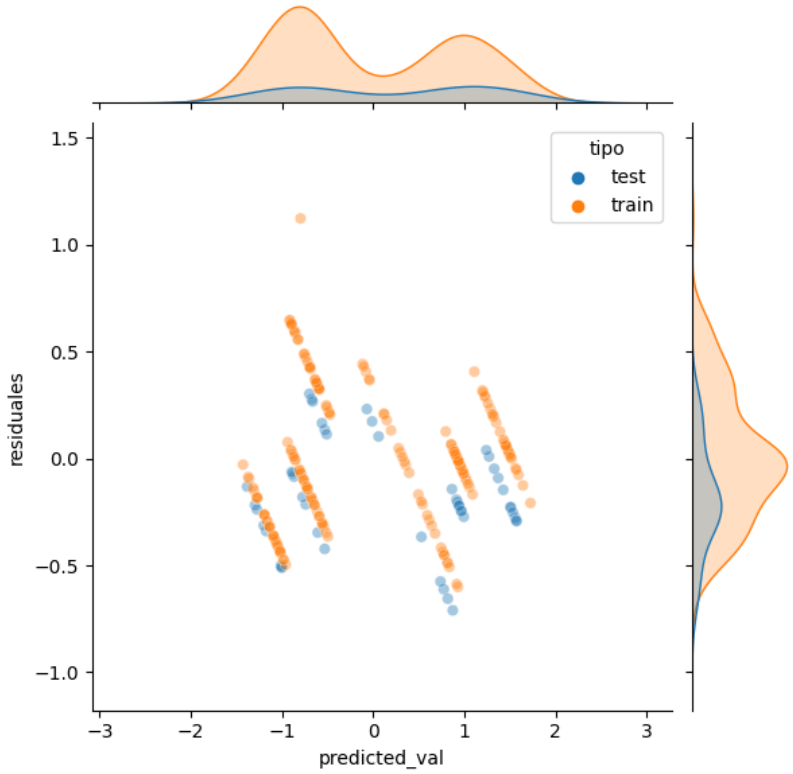

Figura 11: Gráfico de Homocedasticidad

6. Conclusiones

Podemos confirmar que hay dependencia entre las variables de temperatura, color, tipo de estrella y luminosidad puesto que los plots informativos nos dieron los suficientes datos para saberlo.

Además, nuestros datos coinciden con el comportamiento que hay dentro de el diagrama H-R; y, gracias a la regresión lineal, vemos que ese comportamiento se aplicaría a las siguientes series de datos que se le añadieran al modelo, por lo que sería de gran ayuda en el caso de que alguna persona quisiese confirmar por sí misma este comportamiento.

Referencias

[1] The Hertzsprung-Russell diagram - CESAR - Cosmos. (s. f.). https://www.cosmos.esa.int/web/cesar/the-hertzsprung-russell-diagram

[2] Córdova, B. (2022, 31 julio). Rojo, naranja, azul, amarillo: ¿Por qué las estrellas son de diferente color? Enséñame de Ciencia. https://ensedeciencia.com/2022/07/31/rojo-naranja-azul-amarillo-por-que-las-estrellas-son-de-diferente-color/

[3] Star Spectral Classification. (s. f.). http://hyperphysics.phy-astr.gsu.edu/hbasees/Starlog/staspe.html

[4] diagrama de Hertzsprung-Russell — Sociedad española de astronomía. (s. f.). https://www.sea-astronomia.es/glosario/diagrama-de-hertzsprung-russell

[5] Star dataset to predict star types. (2019, 21 octubre). Kaggle. https://www.kaggle.com/datasets/deepu1109/star-dataset

[6] Acerca de la regresión lineal. (s. f.). México — IBM. https://www.ibm.com/mx-es/analytics/learn/linear-regression