DIPLOMATURA EN CIENCIA DE DATOS, APRENDIZAJE AUTOMÁTICO Y SUS APLICACIONES

MENTORÍA M03-2023

Descifrando el Universo: apariencia de las galaxias

Directora: Ingrid Vanessa Daza Perilla

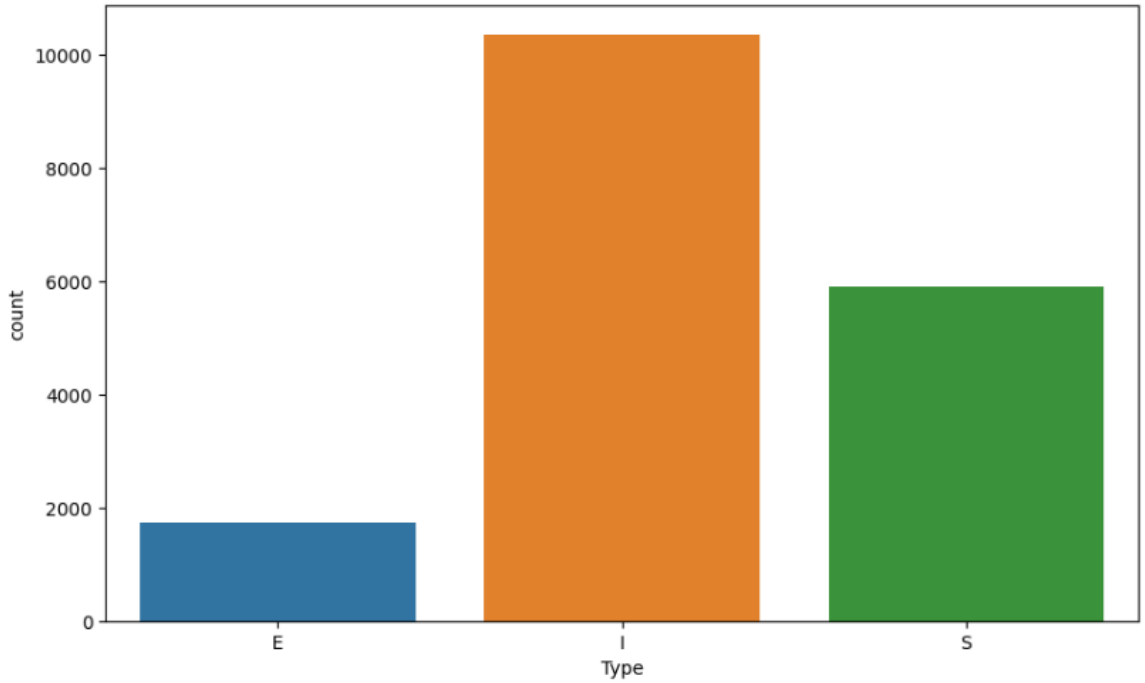
Grupo 2: Ailín Asís, Joaquín Gamalerio y Pablo Velez.

3er Entregable Segunda Parte: Regresión

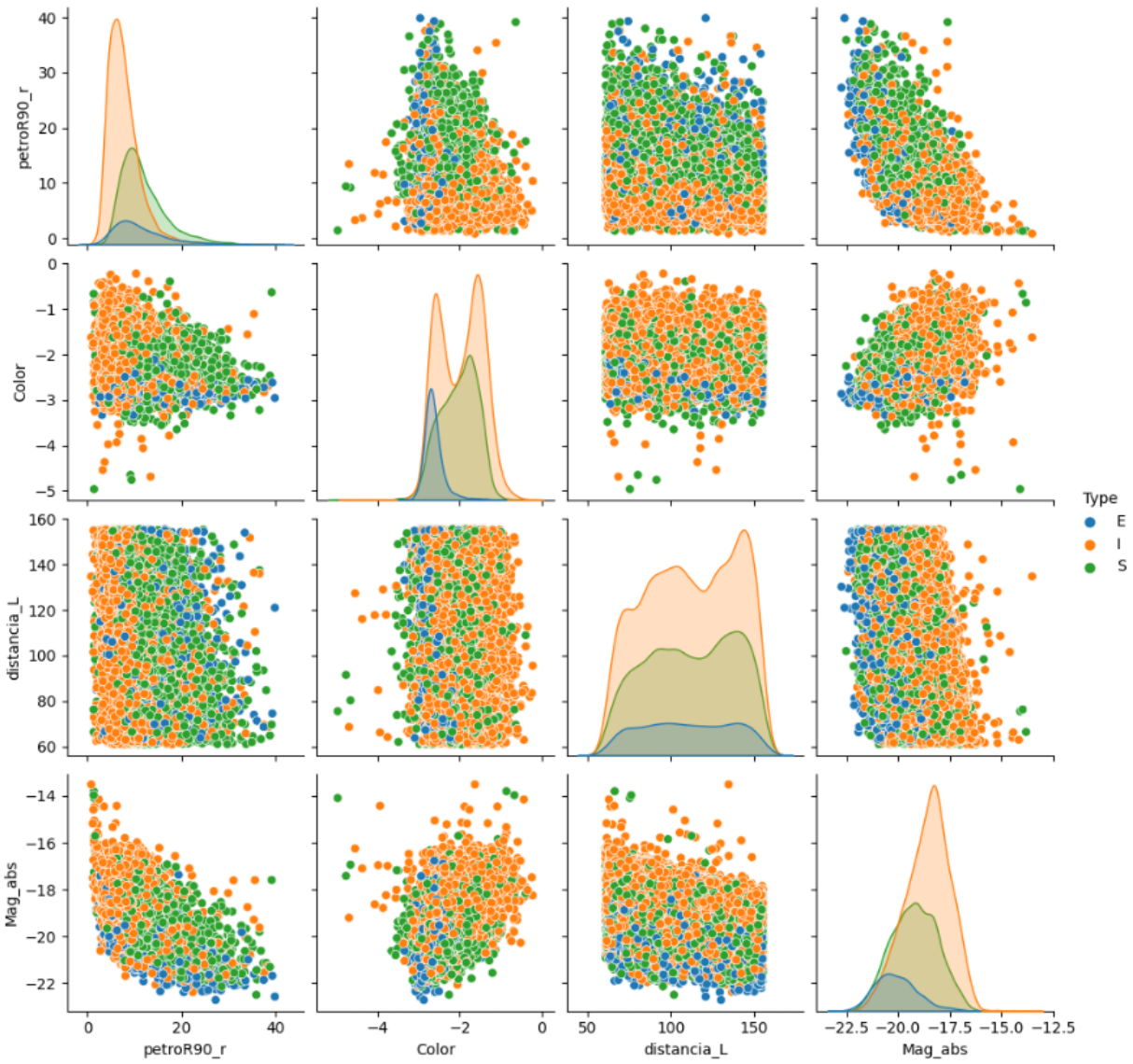
Se exploraron el conjunto de datos obtenidos a partir del **S**loan **D**igital **S**ky **S**urvey (https://skyserver.sdss.org/CasJobs/).

1- Preparación de los datos

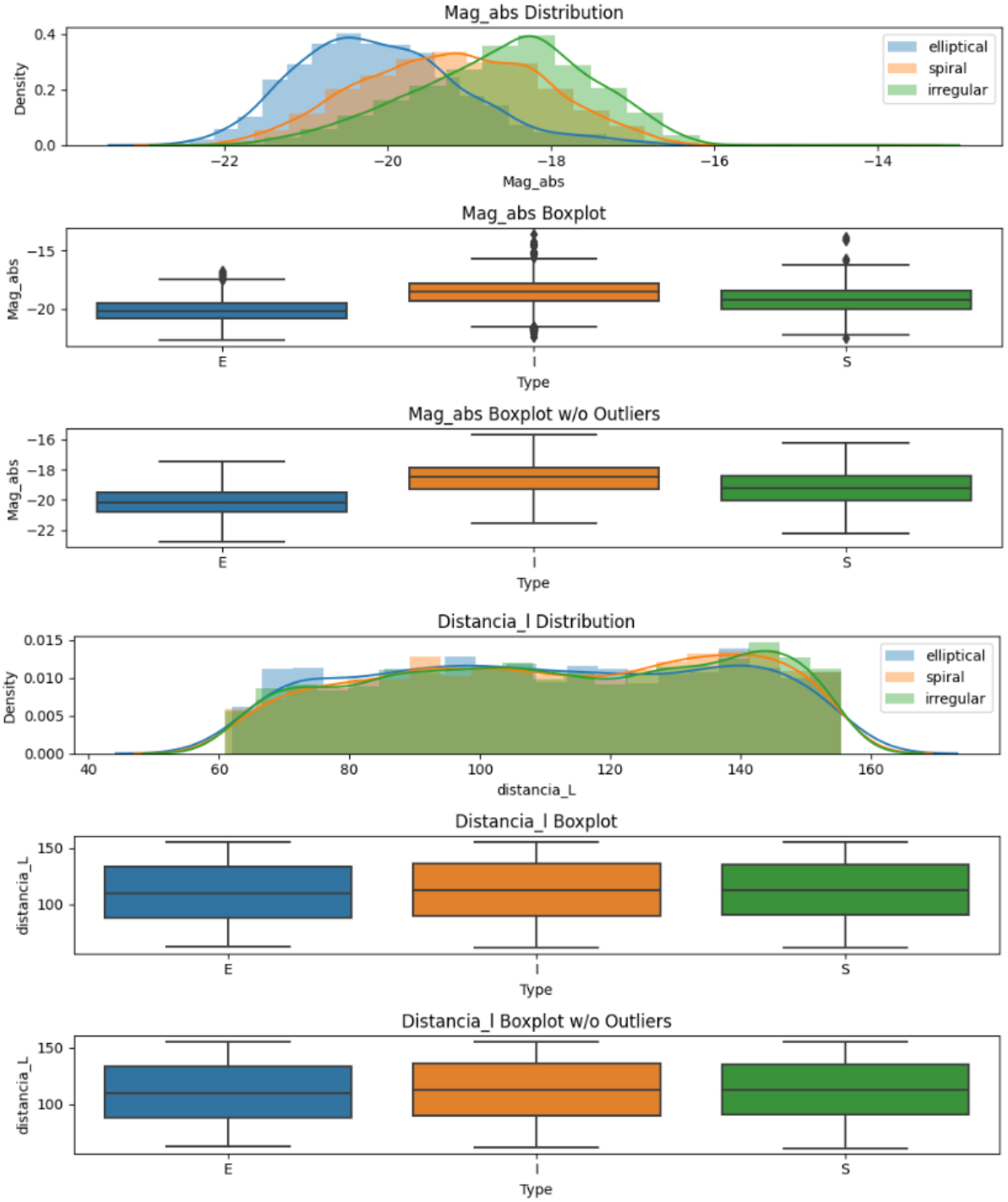
El dataset original es “galaxias_2.csv”. Consta de 32623 registros y 14 variables y no posee valores faltantes.
Las variables son: ‘objID’, 'ra', 'dec', 'modelMag_u', 'modelMag_g', 'modelMag_r', 'modelMag_i',
'modelMag_z', 'petroR90_r', 'distancia_L', 'Color' ,'elliptical' , 'spiral' , `'uncertain'` y ‘Mag_abs’.
Al dataset final se le eliminaron los duplicados y se le añadió una columna ‘Type’ que especifica el tipo morfológico de cada galaxia, resultando en un dataset de dimensiones (18007, 15).
La distribución de clases en el dataset se presenta a continuación, junto con un pairplot.



Se observa que la clase elíptica es la menos representada mientras que las clases irregulares y espirales están más balanceadas, esto no debería ser un problema si asumimos que el dataset es representativo de la distribución poblacional natural.



Cabe señalar que este dataset contiene dos columnas nuevas; ‘Mag_abs’ y 'distancia_L',
La variable magnitud absoluta 'Mag_abs' es una medida del brillo intrínseco, o cuán brillante aparecería (medido en magnitud aparente) el objeto si estuviera ubicado a una distancia estándar de 10 parsecs (aproximadamente 32.6 años luz) de un observador.
La magnitud absoluta de una galaxia se define midiendo toda la luz radiada por todo el objeto, tratando esa luminosidad integrada como la luminosidad de una fuente puntual o estelar, y calculando la magnitud absoluta de esa fuente puntual tal como se describió.
La variable de distancia ``'distancia_L'` es una medida de la distancia a la galaxia en cuestión, en unidades de megaparsec (3.2 millones de años luz aproximadamente).
A continuación se muestran las distribuciones de valores de ambas variables.

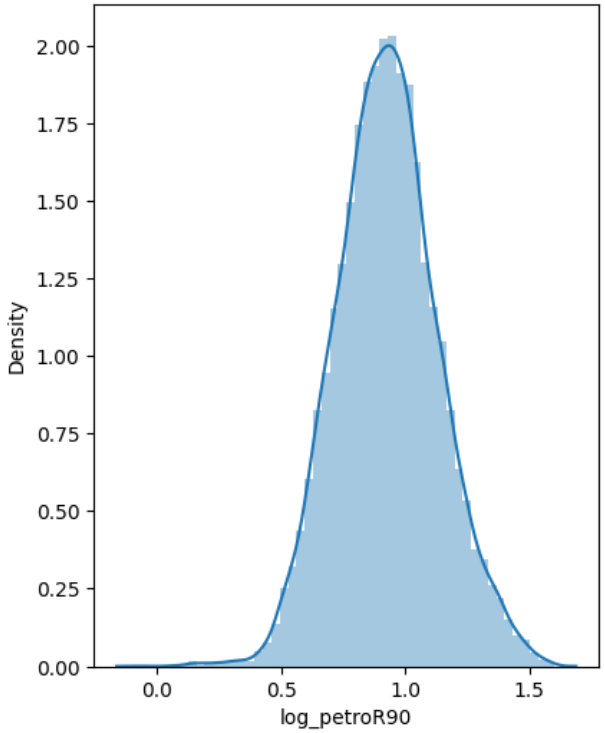


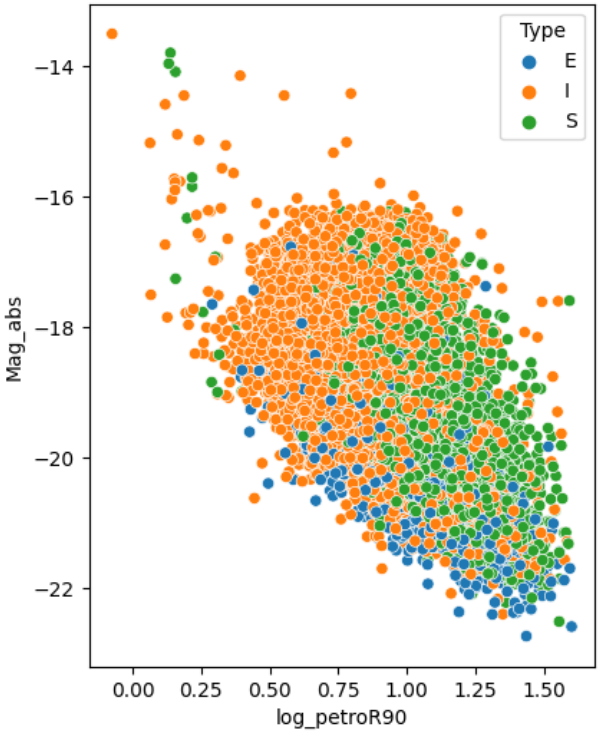
2- **Regresión**

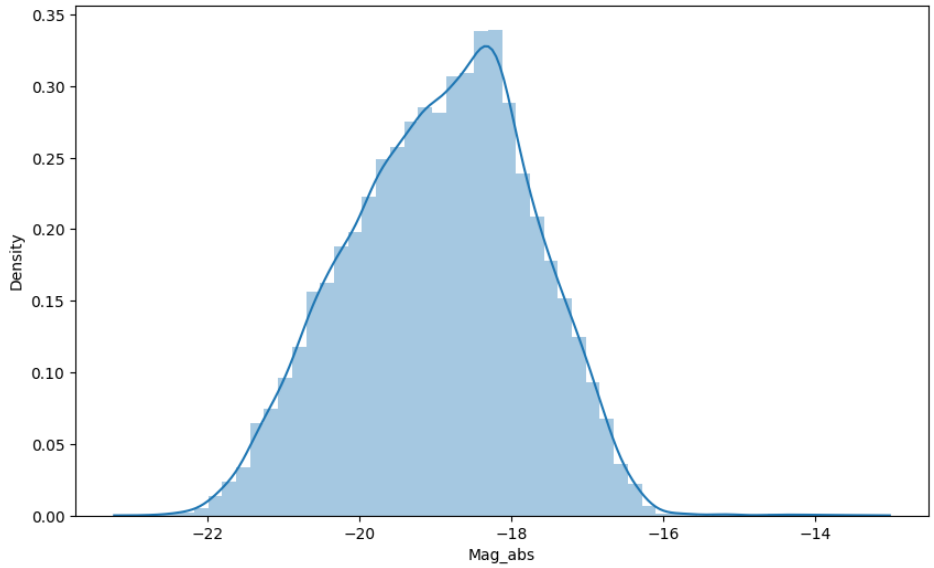
Existe una relación empírica entre el radio efectivo (petro petroR90_r) y la magnitud absoluta para galaxias (datos de Bender et al. 1992, ApJ., 399, 462).
Por lo tanto, el valor a predecir será **Mag_abs** la cual está contenida en un intervalo real de tamaño ~ 9mag, el atributo a usar será el logaritmo en base diez de la variable **petroR90_r**.

Para aplicar el algoritmo de regresión lineal empezamos por dividir al dataset en 4 conjuntos diferentes: **x_train**, **y_train**, **x_test** y **y_test. x_train** y **y_train** son los datos que se usan para realizar el entrenamiento, donde **y_train** son las etiquetas y **x_train** son los features. **x_test** y **y_test** se usan para realizar el test del modelo, donde nuevamente **x_test** son las features o input y **y_test** es la etiqueta real, que sirve para comparar la predicción con el valor esperado. Esa comparación se realiza utilizando métricas como pueden ser la suma residual de cuadrados **RSS**, el error cuadrático medio **MSE**, el error estándar o desviación estándar **SE** o el coeficiente **R-squared**, que determina la proporción de la varianza que es explicada por la variable independiente, en este caso los features.

A continuación se crea una nueva columna **log_petroR90** cuyos valores son el logaritmo en base 10 de la variable **petroR90_r**. Se grafica **Mag_abs vs log_petroR90** para visualizar la dependencia.







Se observa una tendencia decreciente del valor de la magnitud absoluta con respecto al logaritmo en base 10 del tamaño de la galaxia.

Como se explicó anteriormente se realizó la separación del dataset en dos conjuntos train y test, en este caso se siguió la regla 80-20.

2.1- Modelo Lineal con y sin Regularización

LinearRegression() es una regresión normal sin penalización (regularización), donde se calculan los coeficientes del modelo minimizando la función de costo RSS (Residual Sum of Squares), la solución es analítica y es exacta. Además para nuestro caso, RSS tiene unidades de magnitud absoluta al cuadrado, y es la suma de las diferencias al cuadrado del valor predicho por el modelo y el valor real de etiqueta.

$$RSS = \sum_{i=1}^{n} (y^i - f(x_i))^2$$

Lasso() es un modelo de regresión lineal que incluye regularizaciones, añadiendo un término extra, $\lambda||\beta||$, al RSE llamado elemento de regularización L1. El objetivo de Lasso() es evitar que los pesos crezcan demasiado, buscando evitar overfitting. Además, este modelo no tiene solución analítica cerrada ya que no es diferenciable y recurre a métodos numéricos como descenso por el gradiente para encontrar el mínimo de la función de costo.

Se realizó regresión lineal utilizando ambos modelos y se compararon los resultados.

```
Lasso
Lasso(alpha=0.01, random_state=4212)
```

Como se observa en la imagen anterior el alpha usado (coeficiente lambda del término L1) fue de 0.01. No se probaron otros valores.

Resultados del test:

Linear Regression:

- R2 Score: 0.27899199700100696
- MSE: 1.0034302380303863
- RSE: 3614.3557173854515

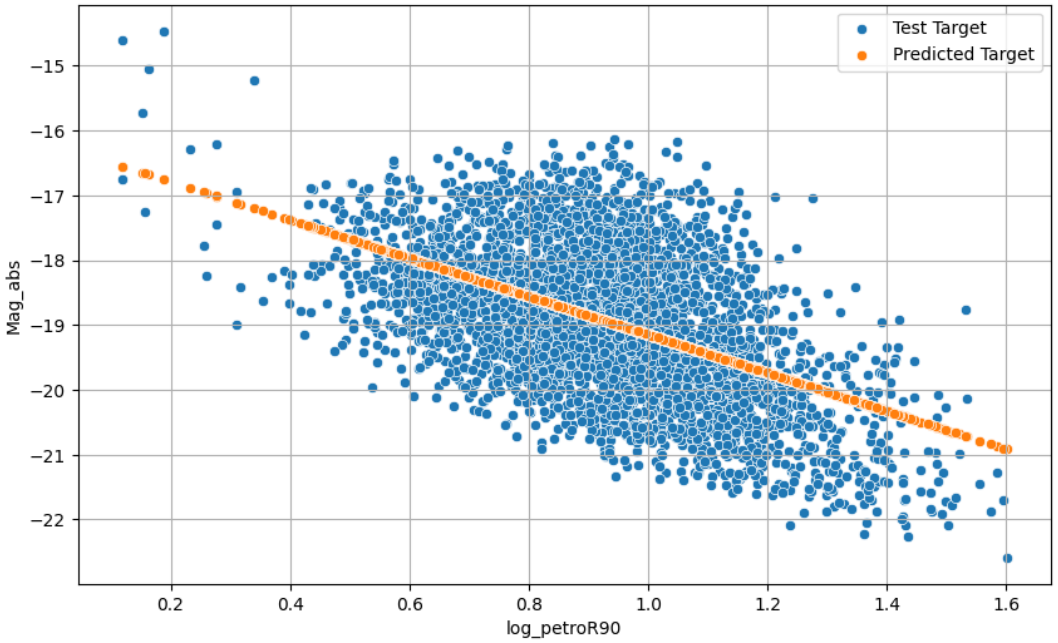
Linear Regression with Lasso Regularization:

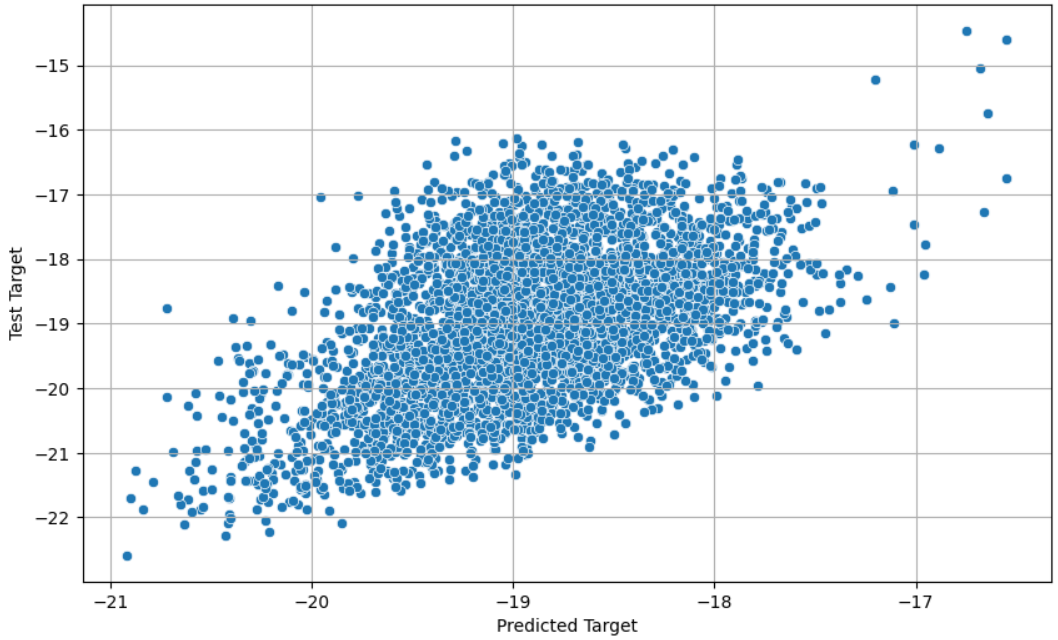
- R2 Score: 0.27612246786614336
- MSE: 1.0074237752600075
- RSE: 3628.7404384865467

Se observa que la regresión lineal normal es más efectiva que la versión regularizada pero por muy poco. Además R2 score tiene un valor relativamente chico, si existe una relación lineal se espera que R2 tenga valores cercanos a 1. Esto quiere decir que la relación no es estrictamente lineal y solo un 0.279% aproximadamente de la varianza es explicada por **log_petroR90**. Se puede decir además que la regularización no es necesaria.

De nuevo se tiene que destacar que los resultados de Lasso varían dependiendo de la semilla utilizada para inicializar los pesos, y puede ser que una semilla diferente nos de un mejor modelo. Para eso se deben inicializar varios modelos con diferentes semillas y elegir el mejor.

A continuación se muestran más gráficos para observar los resultados.





En la segunda imagen se observa la distribución de valores reales vs predichos, idealmente los puntos deberían estar ubicados en una recta y=x, lo cual no se cumple normalmente pero que los puntos siguen una tendencia y=x es lo esperable y eso si se cumple en nuestro modelo.

2.1- Modelo distinguiendo por tipo de galaxia

Se repite el procedimiento pero esta vez distinguiendo por tipo morfológico de galaxia y utilizando LinearRegression().

Se encontraron los siguientes resultados:

Elliptical:

- R2 Score: 0.4452439100100508
- MSE: 0.632555350579261
- Coeficiente: -3.32708517

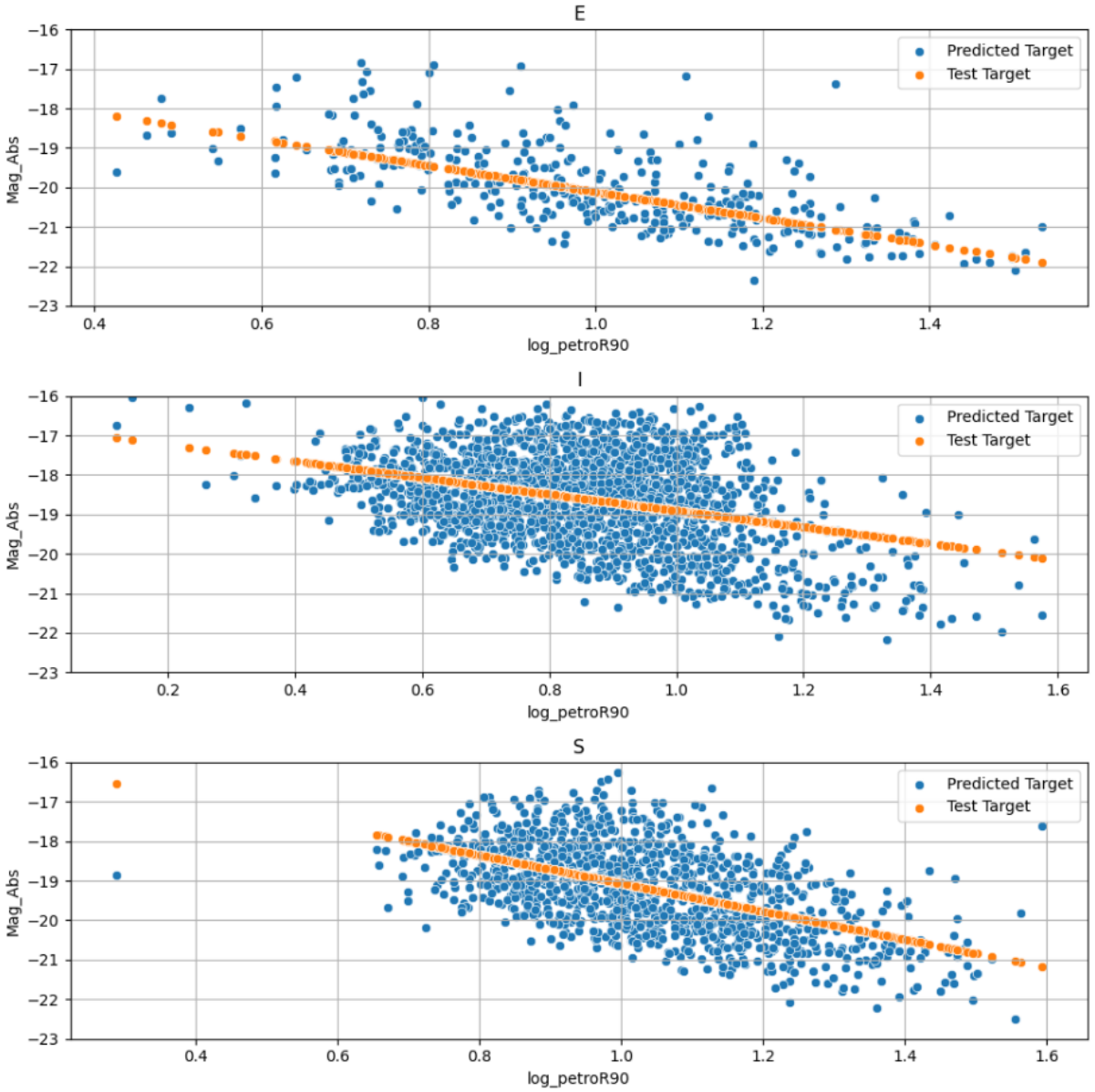
Irregular:

- R2 Score: 0.14991644057874087
- MSE: 1.038591271711517
- Coeficiente: -2.08425229

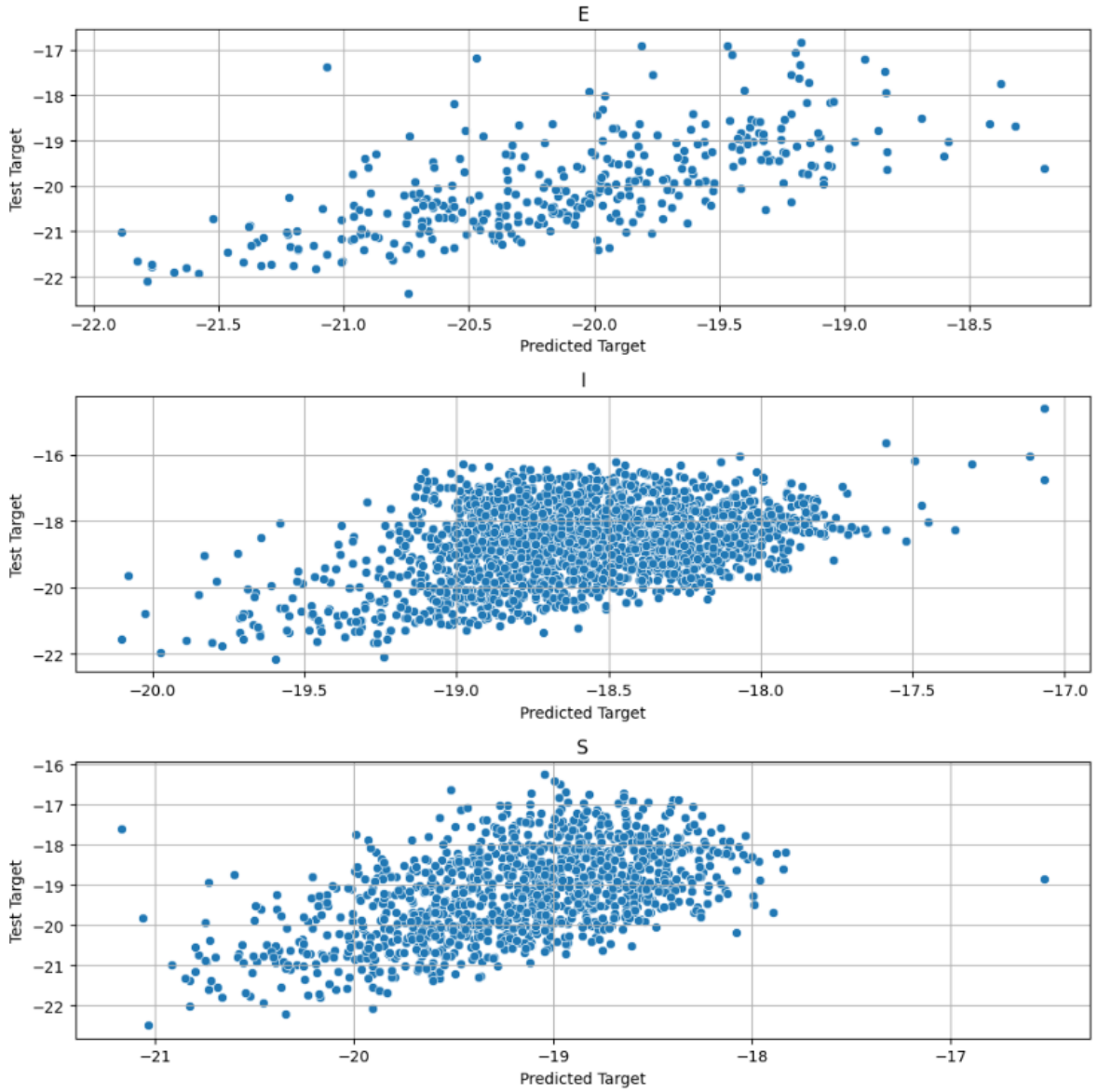
Spiral:

- R2 Score: 0.2861847012598092
- MSE: 0.8544626670983083
- Coeficiente: -3.5546626

Se observa una mejora significativa del R2 score para las galaxias elípticas. Para las galaxias irregulares R2 empeoró y para las galaxias espirales se mantuvo aproximadamente constante.

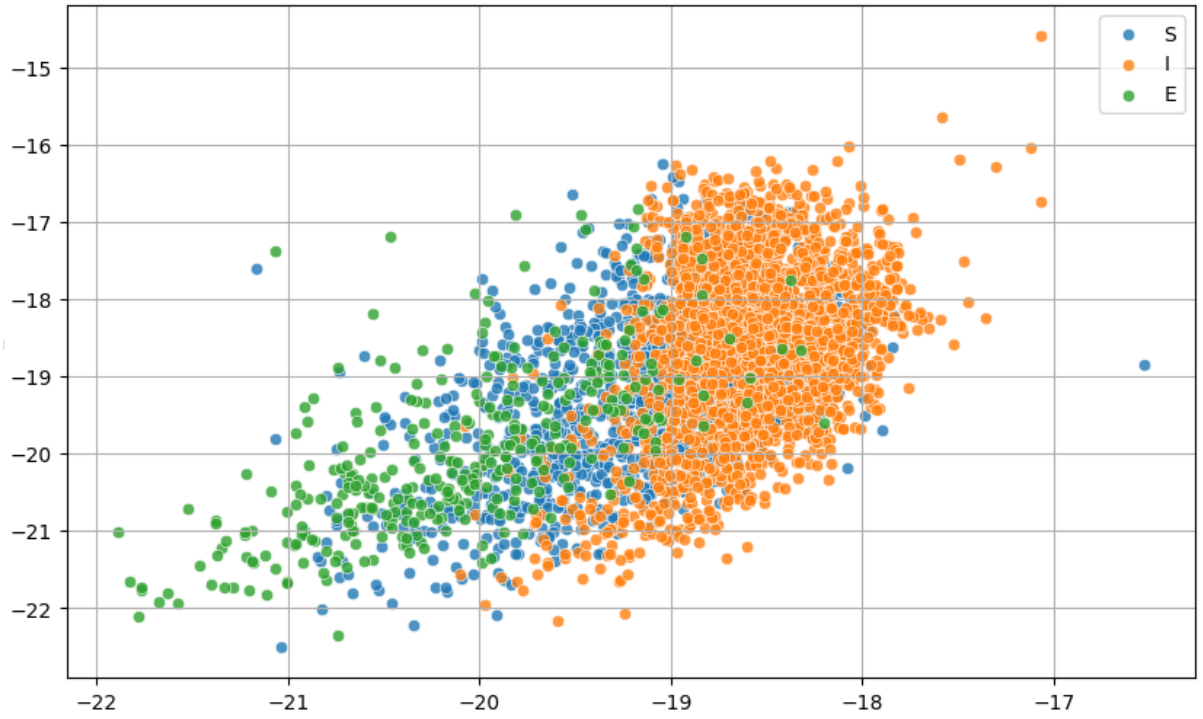


Como se puede observar en la imagen anterior, las galaxias elípticas son las que siguen una tendencia más lineal que las otras.



Se observa en la imagen anterior que nuevamente el valor real vs valor predicho siguen una tendencia y=x, notar que las galaxias Elípticas son las que mejor respetan esta regla.

En la siguiente gráfica se juntan los tres scatter plot anteriores para visualizar mejor lo explicado.



5- Conclusiones

Se realizó Regresión Lineal con y sin regularización sobre el *dataset*: “galaxias_2.csv”, aplicando una limpieza de duplicados y agregando una columna de target.

Se encontró que la regularización L1 o Lasso() no es necesaria, ya que la regresión lineal sin regularización es óptima para problemas con muchos datos, mientras que su contraparte funciona mejor cuando no se dispone de demasiados datos. Además un modelo lineal no es algo complejo que necesite regularización, recordar que lo que se busca al regularizar es disminuir la complejidad del modelo.

Luego de realizado el proceso anterior, se buscó repetirlo distinguiendo por galaxia, encontrando una mejora significativa para R2 de las galaxias elípticas, para las galaxias irregulares empeoró y para las espirales se mantuvo constante.

Como consecuencia, dados los valores pequeños de R2, afirmamos que a grandes rasgos no existe una relación estrictamente lineal entre la Magnitud Absoluta de una galaxia y el logaritmo en base 10 de su tamaño, más bien existe una tendencia lineal que puede ser modelada en su mejor versión con R2 = 0.4452439100100508 para el caso de las galaxias elípticas, es decir, un 40% de la varianza puede ser explicado por log(petroR90_r).