Clasificación de estrellas

Sebastian C. Mariño M, *marino_sebastian@javeriana.edu.co*
Bogotá D.C, Pontificia Universidad Javeriana

***Resumen*—El presente documento tiene como objetivo enumerar los resultados y la metodología usada para establecer un sistema de clasificación de estrellas, usando sus características físicas, mediante el uso de algoritmos de machine Learning como lo pueden ser Knn, Regresión logística o SVM**

I. Business Understanding

Se quiere lograr una clasificación inmediata, en donde a partir de unas características establecidas se pueda identificar rápidamente que clase de estrella es la que se esta analizando, esto con el fin de reducir los tiempos de comparación para empezar la investigación de una nueva estrella.

Es importante conocer el sistema de clasificación moderno, este se le llama clasificación Morgan-Keenan (MK). A cada estrella se le asigna una clase espectral de la antigua clasificación espectral de Harvard y una clase de luminosidad utilizando números romanos como se explica a continuación, formando el tipo espectral de la estrella.

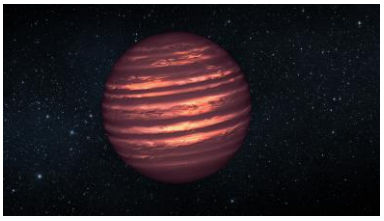
II. Data Understanding

La base de datos usada para este programa se encuentra en: https://www.kaggle.com/brsdincer/star-type-classification

En esta base de datos se nos muestra 6 tipos de estrellas, es decir 6 tipos, estos son:

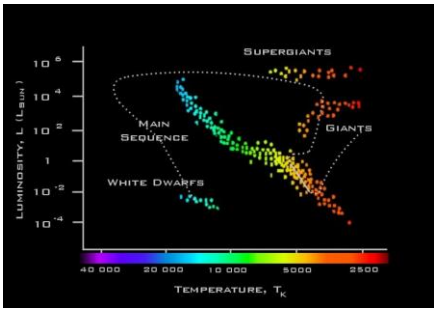
1. Enana Roja: Las enanas rojas son estrellas de muy baja masa, inferior al 40% de la masa del Sol. Su temperatura interior es relativamente baja y la energía se genera a un ritmo lento por la fusión nuclear de hidrógeno en helio a través de la cadena protón-protón. Por consiguiente, estas estrellas emiten poca luz.

2. Enana Marrón: Junto con los planetas, son objetos subestelares. Técnicamente, en el interior de un objeto subestelar no se consume hidrógeno de forma estable, al contrario de lo que ocurre en las estrellas de la secuencia principal. Sin embargo, las enanas marrones sí que consumen deuterio (un isótopo pesado del hidrógeno), al contrario de lo que ocurre en los planetas.



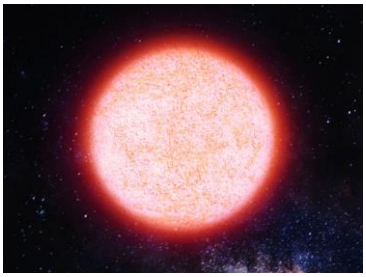
3. Enana Blanca: Estas viejas estrellas son increíblemente densas. Una cucharadita de su materia pesaría en la Tierra tanto como un elefante de 5,5 toneladas. Las enanas blancas tienen típicamente una centésima parte del radio solar, pero su masa es aproximadamente la misma. Las estrellas como el Sol fusionan hidrógeno a helio en sus núcleos. Las enanas blancas son estrellas que agotaron todo el hidrógeno que utilizaban como combustible nuclear.

4. Secuencia Principal: Las estrellas de secuencia principal, también llamadas estrellas enanas, son estrellas que fusionan hidrógeno en sus núcleos. Son enanas porque son más pequeñas que las estrellas gigantes, pero no son necesariamente menos luminosa. Las estrellas de secuencia principal pertenecen a la clase de luminosidad V.



[*] Inteligencia Artificial 2021.

5. Super Gigantes: Las estrellas supergigantes son estrellas con masas comprendidas entre 10 y 50 masas solares y enormes dimensiones, que en el caso de las supergigantes rojas pueden ser del orden de 1000 veces la del Sol. Debido a su gran masa, consumen energía a un ritmo muy elevado, siendo muy luminosas.



6. Hiper Gigantes: Una hipergigante es una estrella excepcionalmente grande y masiva, incluso mayor que una supergigante. Su masa puede ser de hasta 100 veces la masa de nuestro Sol, próxima al límite máximo teórico, el cual establece que la cantidad de masa en una estrella no puede exceder las 120 M⊙ (masas solares).

Ahora se enumerarán las características de las estrellas con las cuales se definen los tipos que se mencionaron anteriormente:

a. Temperatura: Esta variable se mide en grados Kelvin.

b. Luminosidad Relativa: Esta luminosidad es relativa a la luminosidad del sol, por ejemplo, la estrella próxima Centauri tiene una luminosidad relativa de 0,000138, esta se mide en luminosidad solar($L_\odot$)

c. Radio Relativo: Del mismo modo de la luminosidad, este es el radio de la estrella relativo al radio del sol, es decir el sol es la referencia.

d. Magnitud absoluta: Esta es la magnitud aparente ‘m’ que tendría un objeto si estuviera a una distancia de 10 párseces en un espacio completamente vacío sin absorción.

e. Color: Como su nombre lo indica, es el color observable y aparente que tienen las estrellas .

f. Clase espectral: Se dividen de las letras O a la M, Físicamente, las clases indican la temperatura de la atmósfera de la estrella y normalmente se clasifican de más caliente a más fría.

III. Data Preparation

Al conocer los datos se analiza la base de datos, por tanto, se limpia y se transforma en una nueva, en este caso no había características sin datos(NaN), por tanto, este problema no fue necesario de tratar, a continuación, enumerare las características que sufrieron cambios:

- Color: En esta característica se presentaros diferentes problemas, el primero consistía en las mayúsculas ya que se encontraron colores iguales pero al estar uno en mayúsculas y el otro minúsculas se detectaban diferentes, por esto, esta columna se dejó en mayúscula, y se eliminaron los caracteres especiales como ”-”, como segundo problema se notó que habían colores iguales pero escritos de diferentes maneras, por ejemplo “BLUE WHITE” y “WHITE BLUE”, esto se tuvo que corregir, dando como resultado los siguientes 11 colores: 'RED', 'BLUE WHITE', 'WHITE', 'YELLOWISH WHITE', 'PALE YELLOW ORANGE', 'BLUE', 'WHITISH', 'YELLOW WHITE', 'ORANGE', 'YELLOWISH', 'ORANGE RED', por último se reemplazaron caracteres por número, 0 ´RED’, 1 BLUE WHITE’ y así sucesivamente.

- Clase espectral: En esta columna se identificaron 7 clases, como se mencionó en “DATA UNDESTANDING”, estos son: 'M', 'B', 'A', 'F', 'O', 'K', 'G', estos caracteres se reemplazaron por números, ‘M’ por 0, ‘b’ por 1 y así sucesivamente.

- Tipos: Se identificaron números del 0 al 5, los cuales se representan así:

  Red Dwarf - 0

  Brown Dwarf - 1

  White Dwarf - 2

  Main Sequence - 3

  Super Giants - 4

  Hyper Giants – 5

Al final todos estos datos se acoplaron y se genero una base de datos limpia, como se muestra a continuación:

	Temperature	L	R	A_M	Color	Spectral_Class	Type
0	3068	0.002400	0.1700	16.12	0	0	0
1	3042	0.000500	0.1542	16.60	0	0	0
2	2600	0.000300	0.1020	18.70	0	0	0
3	2800	0.000200	0.1600	16.65	0	0	0
4	1939	0.000138	0.1030	20.06	0	0	0
..	...	...	...	...	...	...	...
235	38940	374830.000000	1356.0000	-9.93	5	4	5
236	30839	834042.000000	1194.0000	-10.63	5	4	5
237	8829	537493.000000	1423.0000	-10.73	2	2	5
238	9235	404940.000000	1112.0000	-11.23	2	2	5
239	37882	294903.000000	1783.0000	-7.80	5	4	5

IV. Modelado

Para el modelado se dividí la metodología en diferentes pasos:

- División datos: Se dividió la base de datos en 2, en las características(X) y la ultima columna, los tipos(y), al tenerlos divididos se usó el método de Sklearn

"train_test_split", para dividir los datos en train y test, es decir unos para entrenar el algoritmo deseado y para probarlo. El porcentaje de datos de prueba se vario para encontrar el mejor modelo.

- Normalización: En este punto los datos X tanto de train como de prueba se normalizaron entre 0 y 1 mediante el método "StandardScaler" de Sklearn

- Reducción dimensional: En este paso se redujeron las características mediante PCA usando el método "PCA" de Sklearn, se escogieron 2 componentes en principio los cuales representaban el 78% de la varianza, este valor de componentes se varío entre 2 y 4 se compararon los resultados como se mostrará más adelante.

- Knn: Un modelo escogido, fue el algoritmo de clasificación Knn, para encontrar el mejor modelo, se variaron los hiper-parametros, del k y de la distancia, estas distancias eran: 'manhattan', 'chebyshev' y 'Minkowsky'

- Regresión logística: Este fue el segundo modelo escogido, en donde se asumió X+X^2+1 como hipótesis.

V. Evaluación

Se va a dividir la evaluación con diferentes los mejores resultados de diferentes casos y se mostrara el resultado de sus métodos de evaluación, este caso F1 y MCC. Se hará una tabla resumen de lo encontrado, esto se realizo dividiendo los datos aleatoriamente

Knn

K	Distan	PCA(Var)	% Test	MCC	F1
1	manhattan	3(0.87)	25	0.9606	0.966
1	manhattan	6(0.99)	20	1	1
4	manhattan	6(0.99)	60	0.975	0.979
1	chebyshev	3(0.87)	60	0.966	0.972
11	manhattan	3(0.87)	30	1	1
1	manhattan	6(0.99)	30	1	1
11	manhattan	2(0.77)	20	1	1
1	manhattan	6(0.99)	20	1	1
1	manhattan	2(0.77)	25	0.98	0.9833
1	manhattan	6(0.99)	25	1	1

Regresión Logística

PCA(Var)	% Test	MCC	F1
2(0.77)	25	0.8102	0.8333
6(0.99)	25	1	1
2(0.77)	20	0.779	0.8125
6(0.99)	20	1	1
2(0.77)	40	0.6789	0.7083
6(0.99)	40	1	1
3(0.87)	25	0.8437	0.8666
6(0.99)	25	1	1
4(0.95)	25	0.903	0.91666
6(0.99)	25	1	1

En las tablas anteriores podemos ver que Knn se comporta mejor al tener pocos datos de entrenamiento, también si se quiera hacer PCA, es recomendable usar Knn con k igual a 1 y distancia manhattan, pero teniendo en cuenta que se debe tener una cantidad aceptable de datos de entrenamiento. En cuanto a la regresión logística vemos que no se comporta muy bien con PCA por lo que, si el problema a fuerzas requiriera una regresión logística, lo mejor sería no usar PCA.

Otra prueba realizada se hizo con cross validation, debido a que el Data set podría considerarse pequeño, esto se hizo con Knn y máquina de soporte vectorial lineal.

Knn

K	Distan	Pliegues	F1 PROM
1	manhattan	5	0.7125

En este caso, podemos ver que el F1 promedio con entrenamiento Knn no es muy favorable, por lo que se decidió usar otro algoritmo para encontrar mejores resultados.

SVM

C	Tipo	Pliegues	F1 PROM
1	Linear	5	0.98333

El resultado con SVM linear es satisfactorio, ya que el F1 promedio es alto.

VI. Conclusiones

- El mejor resultado se evidencio con una maquina de soporte vectorial lineal, ya que las pruebas se realizaron con cross Validation

Referencias