Star Type Classification NASA Dataset

Lilian Bour – lilianbour0@gmail.com

Contents

I. Introduction 3
II. Data Exploration 4
III. Unsupervised learning 8
A. K means 8
IV. Supervised learning 9
A. KNN 9
B. One vs All 10
C. Naïve Bayes Classifier 11
V. Conclusion 12
VI. References 12

I. Introduction

Stars aren't just big orange balls of fire, there exist multiple kind of stars with many colours, size and temperature as it can be seen on the Figure 1.

Figure 1. Types of stars

A Hertzsprung-Russell diagram is a scatter plot that can be used to show the relationship between various stars characteristics as shown on the Figure 2 and Figure 3.

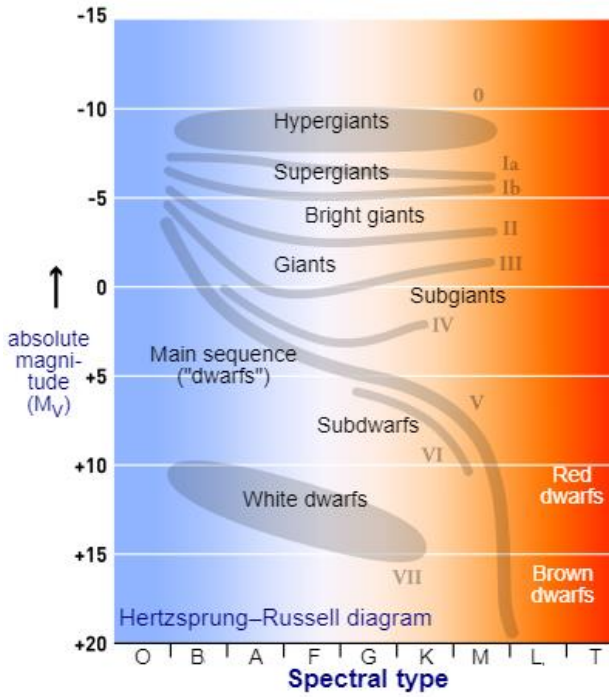

Figure 2. Hertzsprung-Russell diagram 1

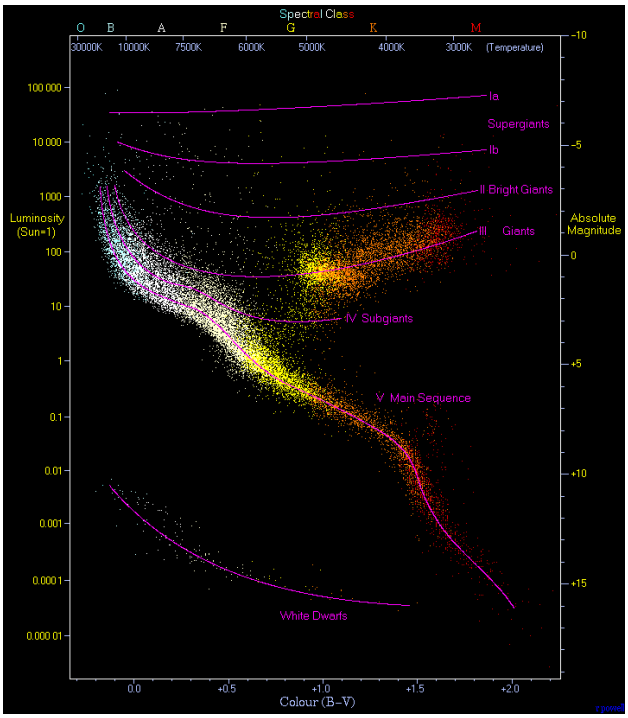

Figure 3. Hertzsprung-Russell diagram 2

II. Data Exploration

The dataset is composed of 240 stars belonging to 6 different kind of stars such as :

- Red Dwarf (**0**)
- Brown Dwarf (**1**)
- White Dwarf (**2**)
- Main sequence (**3**)
- Super Giants (**4**)
- Hyper Giants (**5**)

And the 6 main features are :

- **Temperature** in Kelvin.
- **R⊙**, the solar radius, a unit of distance used to express the size of stars in astronomy relative to the Sun.
- **L⊙**, the luminosity of a given star. Luminosity is an absolute measure of radiated electromagnetic power (light), in this case it's used in the terms of the luminosity of the Sun.
- **Absolute Magnitude**, a measurement of the luminosity of a celestial object. An object's absolute magnitude is equal to the apparent magnitude that the object would have if it were viewed from exactly 32.6 light-years.
- The **colour** of the star.
- The **Spectral class** is a spectral classification based on spectral characteristic obtained via analyse of the electromagnetic radiation.

As we can see on the Figure 4, the distribution is homogeneous, there's 40 stars of each type.

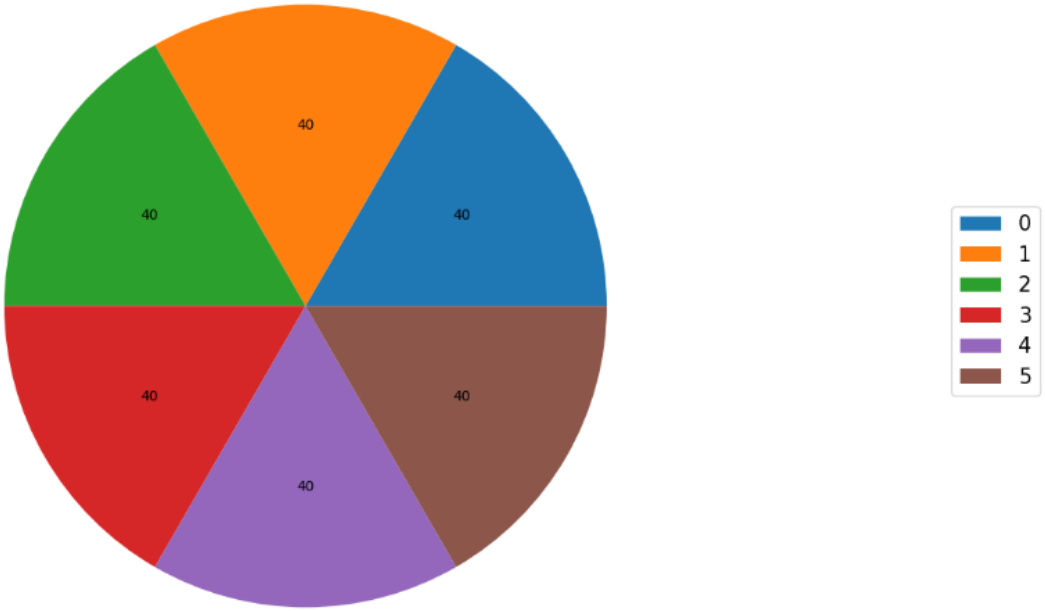

Figure 4. Numbers of stars per type

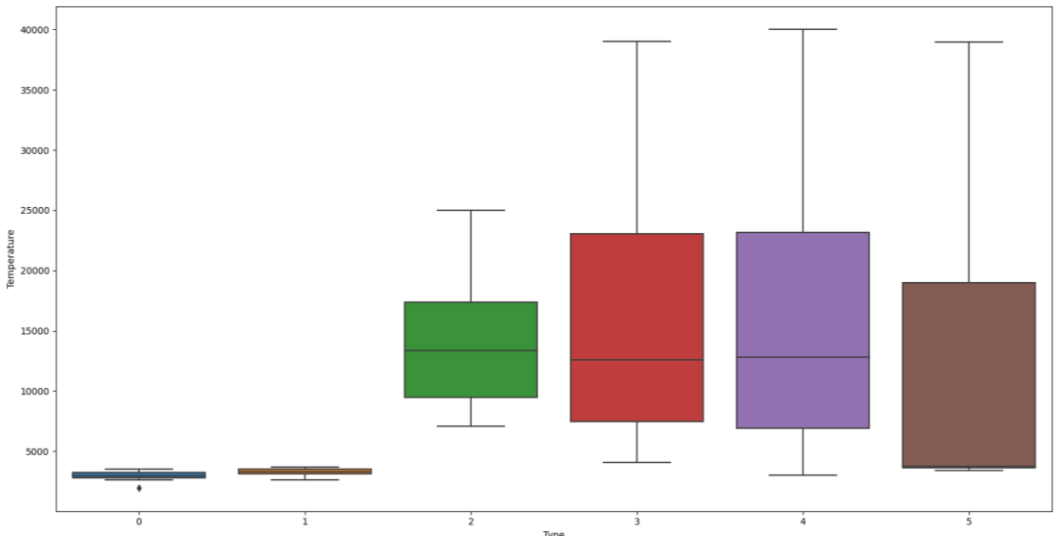

Figure 5. Boxplots per type for the Temperature

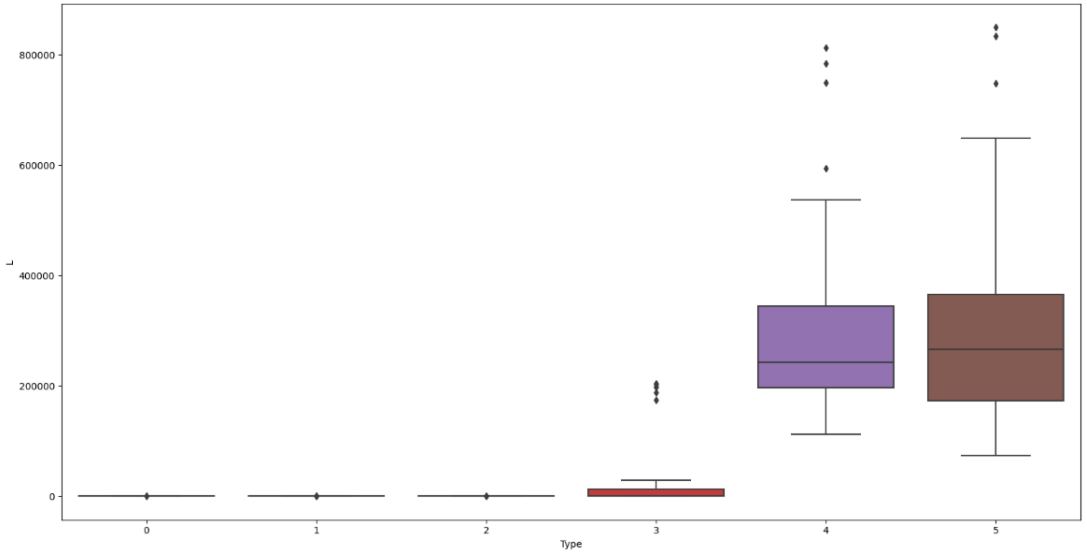

Figure 6. Boxplots per type for the Luminosity

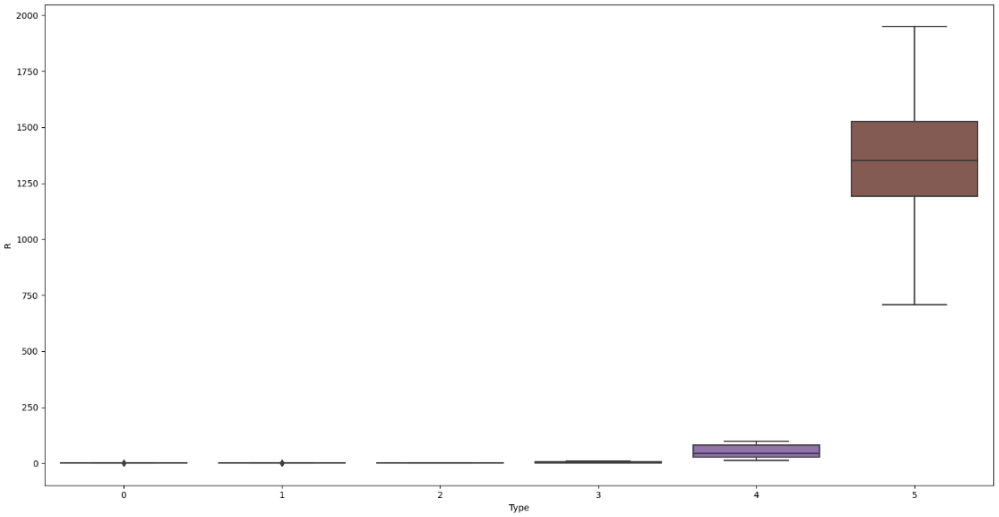

Figure 7. Boxplots per type for the Solar Radius

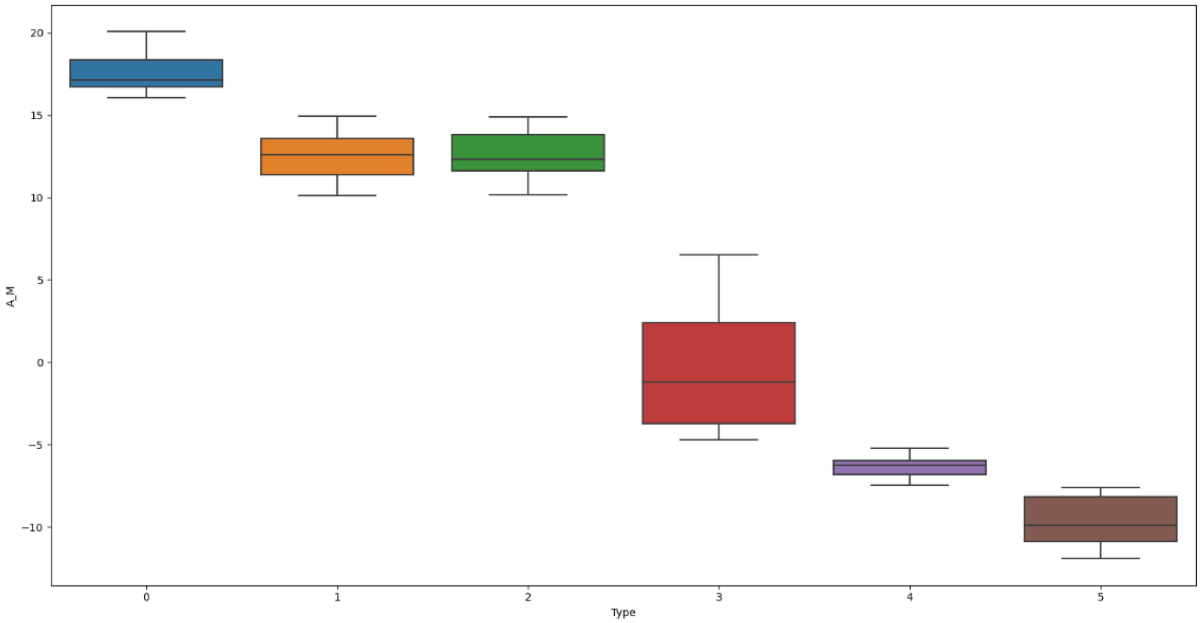

Figure 8. Boxplots per type for the Absolute Magnitude

As we can see, it seems that temperature and absolute magnitude are the feature that separate the star type the most easily.

As shown on the Figure 9, a lot of the colours are written differently multiple times such as “Blue white” and “Blue-White”. After some pre-processing, we can see more clearly on the Figure 10 that most of the stars are Blue, Blue White or Red.

Color	
Blue	56
Blue White	10
Blue white	4
Blue-White	1
Blue-white	26
Orange	2
Orange-Red	1
Pale yellow orange	1
Red	112
White	7
White-Yellow	1
Whitish	2
Yellowish	1
Yellowish White	3
white	3
yellow-white	8
yellowish	2

Figure 9. Coulours before changes

Color	
Blue	56
Blue White	41
Orange	2
Orange Red	1
Pale Yellow Orange	1
Red	112
White	10
White Yellow	9
Whitish	2
Yellowish	3
Yellowish White	3

Figure 10. Coulours after changes

And finally, most of the stars are either of spectral class K, G or B.

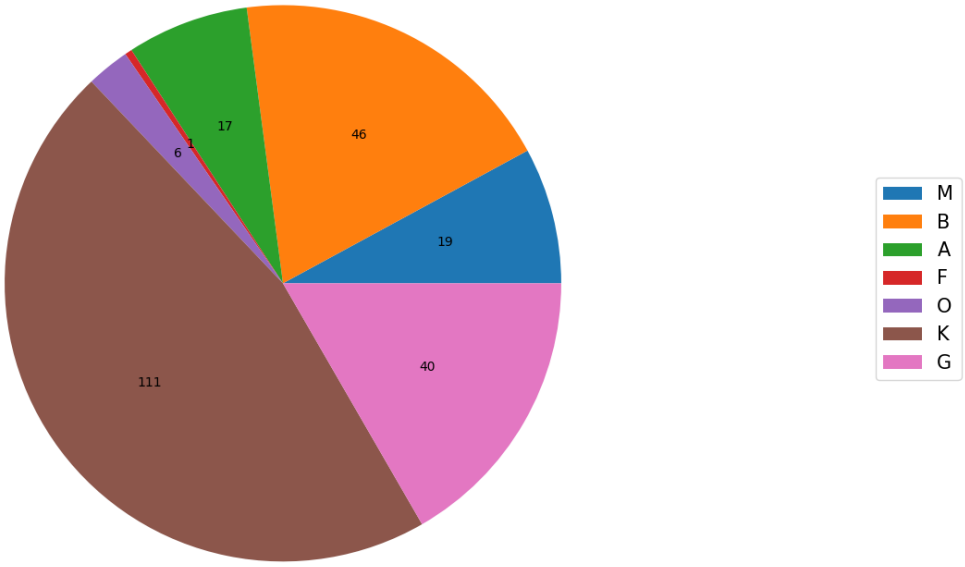

Figure 11. Numbers of Stars per Spectral Class

Using a PCA to plot the data, a graph similar to an Hertzsprung-Russell diagram is obtained with brown and red dwarf near each other (0 and 1), the main sequence in the middle (3), the white dwarfs in the bottom (2) and the giants (4 and 5) at the top.

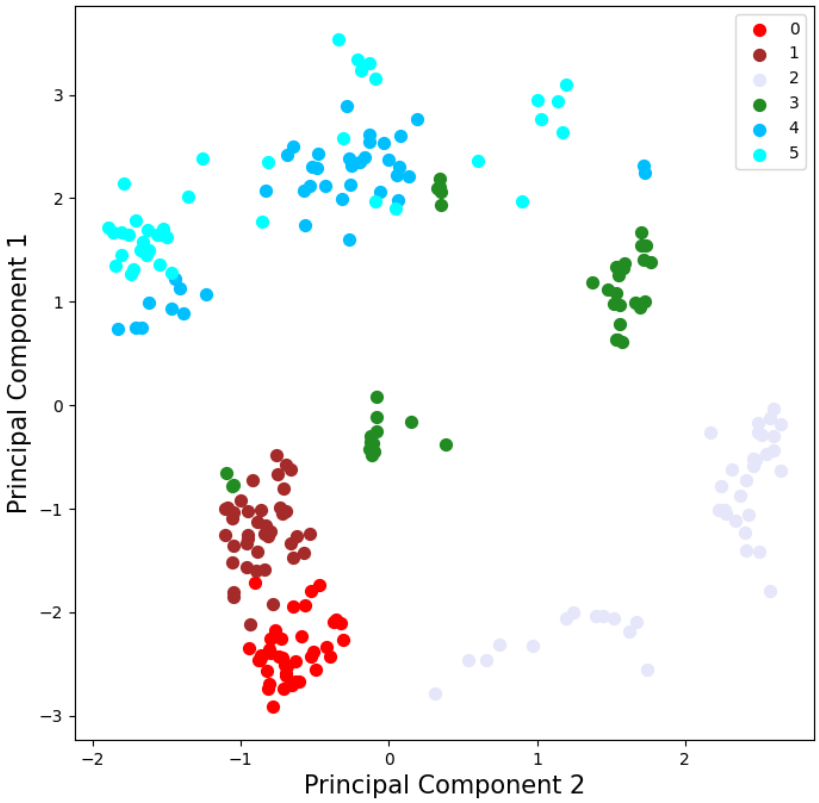

Figure 12. PCA showing the types of stars

Using another dimension reduction method, T-SNE, to plot the data, we can still see some similarities to the Hertzsprung-Russel diagram. The red and brown dwarf (0 and 1) are near each other again. Likewise for the giants (4 and 5). The main sequence (3) is still in the middle and the white dwarfs (2) are at the bottom.

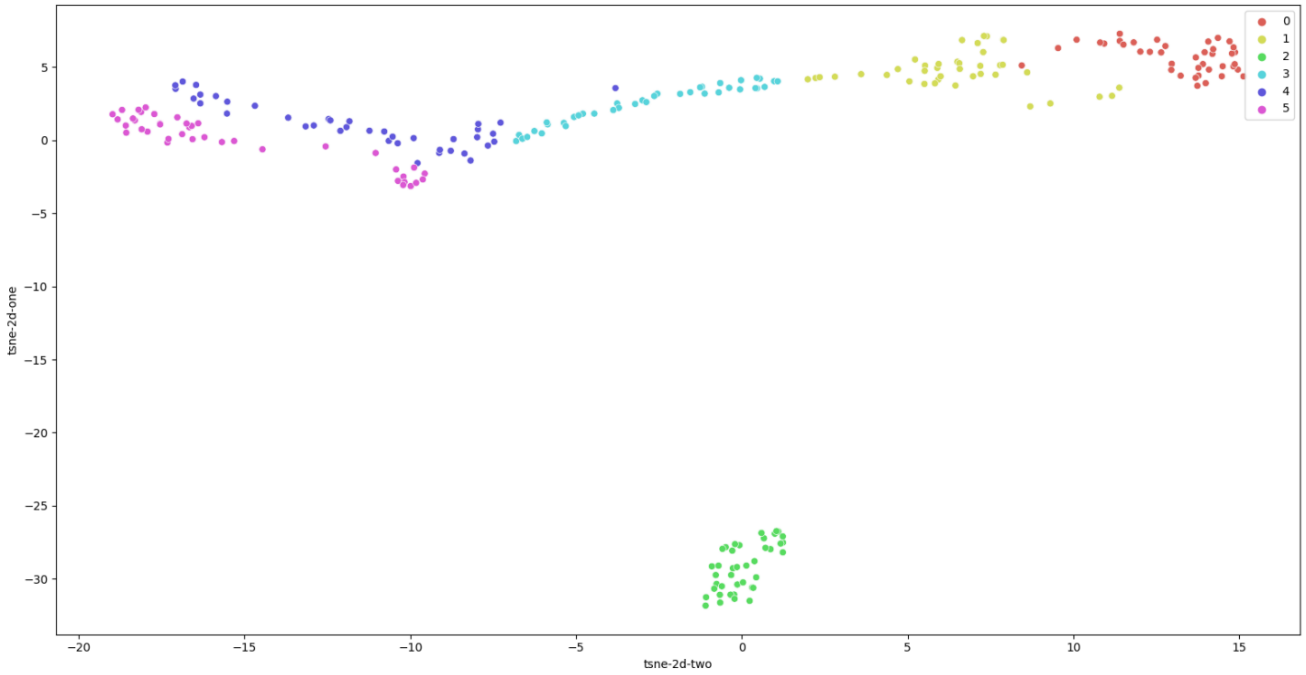

Figure 13. T-SNE showing the types of stars

In our case, the PCA seems better to represent the data in two dimensions, since the representation is mostly a straight line with the T-SNE.

III. Unsupervised learning

A. K means

A K means can be used to see how well it would cluster the types of stars. To begin, the best k, 6 is determined with the elbow method (Figure 14).

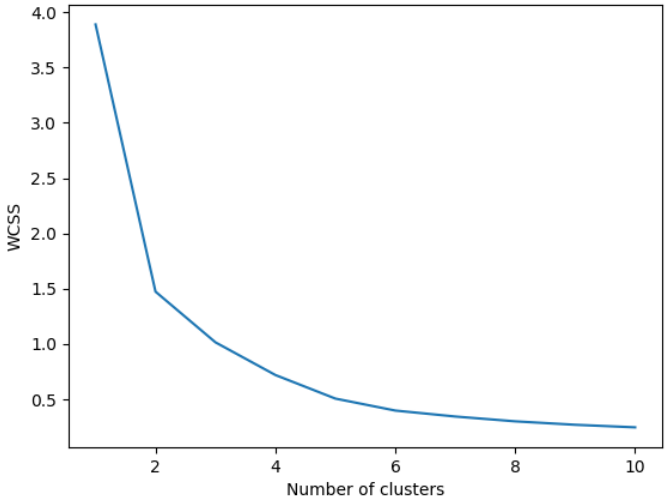

Figure 14. Elbow Method

As expected, the results are not very good since the clusters aren't shaped in circle forms. Indeed, Hyper Giants and Super Giants (cluster 0 and 5) are not well clustered. However, compared to the PCA it is still possible to find some similarities like the red and brown dwarfs (cluster 1 and 3), and for the white dwarfs (cluster 2). The main sequence (cluster 4) is somewhat well clustered but some of it have been clustered with Giants or brown dwarfs.

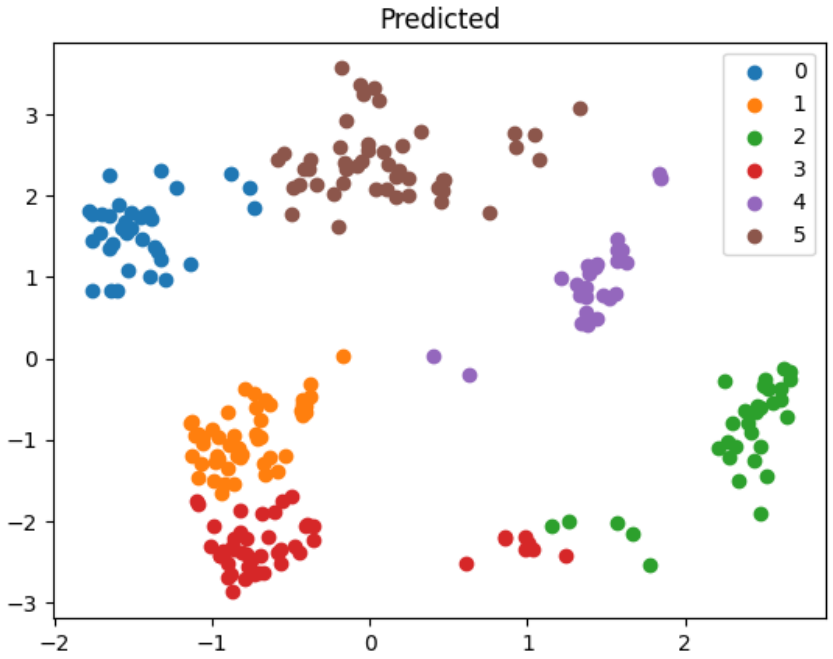

Figure 15. K means clusters

IV. Supervised learning

A. KNN

This dataset can be used to predict the type of start from the 6 features, a KNN can be used with k = 4 as shown on the Figure 16.

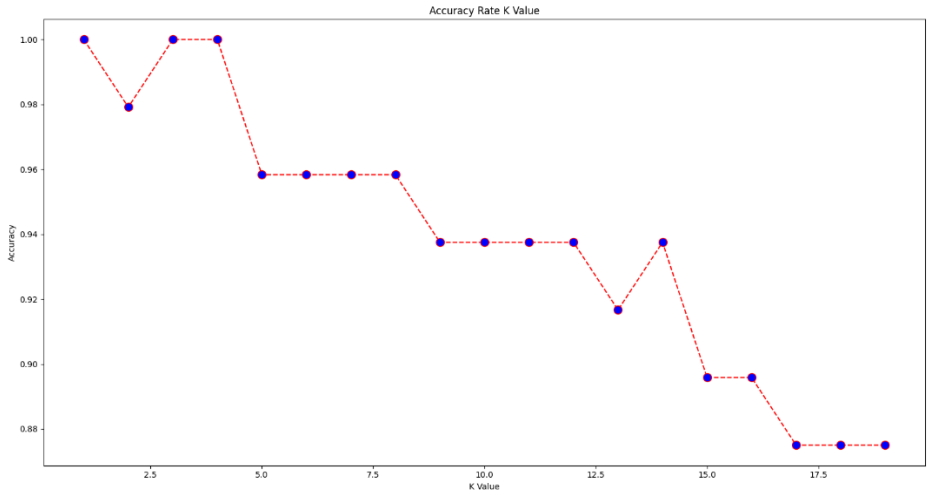

Figure 16. Accuracy rate for each K value

We can see on the confusion matrix (Table 1), that 1 red dwarf is misclassified as a brown dwarf. This is probably due to red and brown dwarf being similar (see Hertzsprung–

Russell diagram) and 1 red dwarf is among brown dwarfs in both the PCA and T-SNE plots. The same phenomenon can be observed for the misclassified hyper giant which is similar to a super giant.

			Actual					
			Red Dwarf	Brown Dwarf	White Dwarf	Main Sequence	Super Giants	Hyper Giants
			0	**1**	**2**	**3**	**4**	**5**
Predicted	Red Dwarf	**0**	12	0	0	0	0	0
	Brown Dwarf	**1**	1	4	0	0	0	0
	White Dwarf	**2**	0	0	11	0	0	0
	Main Sequence	**3**	0	0	0	3	0	0
	Super Giants	**4**	0	0	0	0	7	1
	Hyper Giants	**5**	0	0	0	0	0	6

Table 1. KNN Confusion Matrix

Globally KNN gives some good results (Table 2) with an overall accuracy of 0.96 on a test set. With precision, recall and f1 score being high across the board.

		Precision	**Recall**	**f1 score**
Red Dwarf	**0**	0,92	1,00	0,96
Brown Dwarf	**1**	1,00	0,80	0,89
White Dwarf	**2**	1,00	1,00	1,00
Main Sequence	**3**	1,00	1,00	1,00
Super Giants	**4**	1,00	0,88	0,93
Hyper Giants	**5**	0,86	1,00	0,92
			Accuracy	**0,96**

Table 2. KNN Metrics

B. One vs All

As we can see on the table below, the One vs All algorithm model struggle ton classify Main sequence stars (3), Super Giants (4) and Hyper Giants (5). It could be explained by the fact that those 3 types have some similar features as we can see on the data exploration part. Furthermore, their cluster appears less distinct on the PCA too.

			Actual					
			Red Dwarf	Brown Dwarf	White Dwarf	Main Sequence	Super Giants	Hyper Giants
			0	**1**	**2**	**3**	**4**	**5**
Predicted	Red Dwarf	**0**	12	0	0	0	0	0
	Brown Dwarf	**1**	0	10	0	0	0	0
	White Dwarf	**2**	0	0	12	0	0	0
	Main Sequence	**3**	0	0	0	6	1	0
	Super Giants	**4**	0	1	0	1	3	0
	Hyper Giants	**5**	0	0	0	2	0	7

Table 3. OneVsAll Confusion Matrix

Globally, the results are still good except for the main sequence (3) type with a precision of 0.67 and recall of 0.86. However, the overall accuracy is 0.90 which is good.

		Precision	Recall	f1 score
Red Dwarf	0	1,00	1,00	1,00
Brown Dwarf	1	0,91	1,00	0,95
White Dwarf	2	1,00	1,00	1,00
Main Sequence	3	0,67	0,86	0,75
Super Giants	4	0,75	0,60	0,67
Hyper Giants	5	1,00	0,78	0,88
			Accuracy	0,90

Table 4. OneVsAll Metrics

C. Naïve Bayes Classifier

Here a brown dwarf (0) is classified as a red dwarf (1), the situation is similar to the one with the KNN model. It can be explained by the fact that red and brown dwarf are very alike.

			Actual					
			Red Dwarf	Brown Dwarf	White Dwarf	Main Sequence	Super Giants	Hyper Giants
			0	1	2	3	4	5
Predicted	Red Dwarf	0	6	1	0	0	0	0
	Brown Dwarf	1	0	7	0	0	0	0
	White Dwarf	2	0	0	7	0	0	0
	Main Sequence	3	0	0	0	10	0	0
	Super Giants	4	0	0	0	0	10	0
	Hyper Giants	5	0	0	0	0	0	7

Table 5. Naïve Bayes Classifier Confusion Matrix

The model is nearly perfect with an accuracy of 0.98, the only categories that aren't perfectly predicted aren't because of the misclassification of a brown dwarf.

		Precision	Recall	f1 score
Red Dwarf	0	1,00	0,86	0,92
Brown Dwarf	1	0,88	1,00	0,93
White Dwarf	2	1,00	1,00	1,00
Main Sequence	3	1,00	1,00	1,00
Super Giants	4	1,00	1,00	1,00
Hyper Giants	5	1,00	1,00	1,00
			Accuracy	0,98

Table 6. Naïve Bayes Classifier Metrics

V. Conclusion

Multiple methods to visualise and classify stars have been used. It seems that a PCA is the best way to represent the data, since the visualisation is similar to an Hertzsprung-Russell diagram. The classification is better when using a Naïve bayes classifier with an accuracy near 1 at 0.98.

VI. References

- https://www.kaggle.com/brsdincer/star-type-classification
- https://en.wikipedia.org/wiki/Absolute_magnitude
- https://en.wikipedia.org/wiki/Luminosity
- https://en.wikipedia.org/wiki/Solar_radius
- https://cosmonova.org/different-types-stars-stellar-evolution/