Star Type Classification / NASA Dataset

# Introduction

Diagram

Description automatically generated

Figure 1. Hertzsprung-Russell diagram

# 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 2, the distribution is homogeneous, there’s 40 stars of each type.

Chart, pie chart

Description automatically generated

Figure 2. Numbers of stars per type

Chart, box and whisker chart

Description automatically generated

Figure 3. Boxplots per type for the Temperature

Chart, box and whisker chart

Description automatically generated

Figure 4. Boxplots per type for the Luminosity

Chart

Description automatically generated

Figure 5. Boxplots per type for the Solar Radius

Chart, box and whisker chart

Description automatically generated

Figure 6. 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 7, 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 8 that most of the stars are Blue, Blue White or Red.

Text

Description automatically generated

Figure 7. Coulours before changes

Text

Description automatically generated

Figure 8. Coulours after changes

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

Chart, pie chart

Description automatically generated

Figure 9. 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.

Chart, scatter chart

Description automatically generated

Figure 10. 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.

Chart, scatter chart

Description automatically generated

Figure 11. 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.

# Unsupervised learning (Kmeans, Hierarchical clustering)

# Supervised learning (KNN)

# Conclusion

# References

Data : <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>