**Data:**

For my second dataset I will be exploring this dataset. It's about stars! The metrics are

- Absolute Temperature (in K)
- Relative Luminosity (L/Lo) (brightness) Lo = 3.828 x 10^26 Watts (Avg Luminosity of Sun)
- Relative Radius (R/Ro) Ro = 6.9551 x 10^8 m (Avg Radius of Sun)
- Absolute Magnitude (Mv) (bigness)
- Star Color (white,Red,Blue,Yellow,yellow-orange etc)
- Spectral Class (O,B,A,F,G,K,,M)
- Star Type **(Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence , SuperGiants, HyperGiants)**

**Notes:**

I believe that this dataset works from the **Harvard spectral classification**, as well as the Yerkes spectral classification system for luminosity and class respectively. Typically the Harvard classification system is made up of other variables already present in the dataset, so I am unsure of the efficacy of its inclusion

**Preprocessing:**

This dataset has mostly done its own preprocessing. There were no nulls. I converted class, color, and type to ints so that they would work nicely with the algorithms. I used Pipeline and StandardScaler to normalize my data.

**Data Understanding:**

Some very important notes about this dataset are that the target class, star type, is evenly distributed. This is not actually true in real life, but it's really handy for the data. There are exactly 40 of each kind of star, for zero to five, totaling 240 stars. Spectral class and star color follow something much closer to a logarithmic distribution, with the vast majority of the entries being of the first kind, and the following shrinking by about a factor of two. Temperature, luminosity, radius and absolute magnitude are all quite similar in terms of their distribution. They are all high invariability, and display moderate clustering, visually speaking.

**Analysis And Results:**

All models performed extremely well. So well to the point that relative comparisons are almost pointless. All models performed in the range of 0.94-1.00, with the random forest performing the best and knn performing the worst, but, even counting in the f1 score and the confusion matrix, the models all performed extremely similarly.

Generally though, the classification of stars is based pretty simply on the variables in this dataset, so it is not a surprise that all of the models performed so well. A decision tree performing best makes sense, while most of the other models would probably produce overly

complex overfit models. Typically, in real world classification, the process of actually deciding the class of a planet is essentially a decision tree.

I do wonder if it would be better to entirely abandon convention and focus on allowing for more complex models based on machine learning techniques. It's hard to know but that would be neat. If I were to explore that, I would want a less clean dataset with more variables to play with.

**Sources:**

https://lco.global/spacebook/stars/types-stars/
https://www.youtube.com/watch?v=Y5VU3Mp6abI