**Objectives**

The model will be focused on the interpretation of how stars follow a pattern based on their physical features. The model will help researchers and scientists accurately classify stars based on these physical features.

**Data Description**

The dataset is called “The Stars Dataset” and consists of several features of stars. The purpose of this analysis is to prove that the stars follow a certain graph in the celestial space specifically called Hertzsprung-Russell Diagram, which is shown in **Figure 1**.

Chart, scatter chart

Description automatically generated

**Figure 1:** Hertzsprung-Russell Diagram.

The data has 6 features and 1 target:

1. Absolute Temperature (K)
2. Relative Luminosity (L/Lo)
3. Relative Radius (R/Ro)
4. Absolute Magnitude (Mv)
5. Star Color (white, Red, Blue, Yellow, yellow-orange, etc.)
6. Spectral Class (O,B,A,F,G,K,M)
7. Star Type (Target) (Red Dwarf, Brown Dwarf, White Dwarf, Main Sequence, SuperGiants, HyperGiants)

Brown Dwarf = 0, Red Dwarf = 1, White Dwarf = 2, Main Sequence = 3, Supergiant = 4, Hypergiant = 5

The Luminosity and radius of each star is calculated w.r.t. that of the values of Sun.  
Lo = 3.828 x 10^26 Watts  
Ro = 6.9551 x 10^8 m

The model will use the 6 physical features of the stars to accurately predict the Star Type.

**Data Exploration**

Luckily the data was clean. Each column was viewed using seabron and the dataframe description, info, dtypes and value counts were checked to make sure there are no missing values. **Figure 2** is the pairplot of the data.

Chart

Description automatically generated with low confidence

Figure 2: Pariplot of the full dataset.

Text

Description automatically generated

For Feature Engineering, I used a **StandardScaler** to scale the numerical values for the Temperature, Luminosity, Radius and Absolute Magnitude. **OneHotEncoder** was used to encode the categorical values Star color and Specrtal Class.

**Summary of Training**

Three different classifier models were used for the project: Logistic Regression, K Nearest Neighbors and Random Forest. All three models used the same training and test splits (stratified train-test split with sample size of 0.3). There was no need for cross-validation or hyperparameter tuning, since all the models performed very well (based on several performance metrics) and produced high accuracy results.

**Key Findings**

key findings related to the main objectives include:

* The Stars follow a pattern based on their physical features.
* The pattern is very similar to Hertzsprung-Russell Diagram.
* Random Forest performed great for such classification problem and didn’t take a lot of time for training and prediction.

**Model Recommendation**

Of the classifier models, I recommend the Random Forest model as a final model. Random Forest produced the best results and best fit my needs in terms of accuracy and explainability.

**Suggestions**

Suggestions for next steps in analyzing this data:

* Revising and Logistic Regression and the KNN models and try different hyperparameters to see if better results can be achieved.
* Try all three of the models on a different, bigger dataset and see how it performs.
* Use the different hyperparameter tuning techniques to for the models to get better results.