

# The Stellar Trilogy: An Ontology, HR Diagram and ML Harvard Classification

By

**Arihant Tiwari**

Roll Number: 19050

**Gourav Kumawat**

Roll Number: 19127

Course Instructor

**Dr. Parthiban Srinivasan**

Visiting Faculty



**Indian Institute of Science Education and Research (IISER)  
Bhopal**

Date of Submission

November 7, 2021

## **Abstract**

We did a thorough study of different aspects of Stellar Astronomy and Astrophysics. Using Owl-ready2, we created an ontology that can classify the stars based on their temperature, color and dominant spectrum lines into various spectral classes according to Harvard Spectral Classification Scheme. Using python plotting aids, we created HR diagram plots along with other important plots in stellar astronomy. Using KNN, SVM, and Logistic models, we did a two class classification of stellar dataset with a best accuracy of 88.33 %. Using KNN, we did a seven class classification of stellar dataset with a best accuracy of 58.5 %.

# 1. Introduction

Stellar classification plays a significant role in stellar astronomy. The classification provides us with a lot of information about the properties of the stars. In the earlier days of astronomy, these classifications were done manually. In the recent years, the amount of astronomical data and dimensionality of said data is growing rapidly through more and more ambitious astronomical surveys. Sky surveys like, Sloan Digital Data Survey (SDSS) and Gaia Space Mission provide us with enormous amount of detailed data. The era of data-driven astronomy made it impossible to do manual classification of stars. They created a need for the automated classification of stars using machine learning models.

## 2. Astronomy Background

Stars are massive self luminous astronomical bodies that fuels themselves by the thermonuclear reaction going on in there core. They are one of the most prominent parts of the Astronomy and Astrophysics research. With approximately, 200 billion trillion stars in the observable universe, it is practically impossible to study individual stars. Hence, instead of studying all the stars in the universe, astronomers classified stars based on various classification scheme. The work of the researchers is now restricted to studying these individual star types instead of individual stars.

### 2.1 Harvard Spectral Classification Scheme

This classification is based on the absorption line features in the spectrum of stars. This scheme classify stars into seven different categories:

1. **O Star:** These are blue stars with the surface temperature greater than 25000 K. There spectrum have strong ionized Helium absorption lines. For example: *Zeta Orionis Aa*.
2. **B Star:**These are blue white stars with the surface temperature between 25000 K and 11000 K. There spectrum have strong neutral Helium absorption lines. For example: *Regulus*.
3. **A Star:**These are white stars with the surface temperature between 11000 K and 7500 K. There spectrum have strong neutral Hydrogen absorption lines. For example: *Sirius*.

4. **F Star:** These are yellow white stars with the surface temperature between 7500 K and 6000 K. Their spectrum has strong CaII absorption lines. For example: *Polaris*.
5. **G Star:** These are yellow stars with the surface temperature between 5000 K and 6000 K. Their spectrum has strong CaII absorption lines. For example: *Sun*.
6. **K Star:** These are orange yellow stars with the surface temperature between 3500 and 5000 K. Their spectrum has strong CaII absorption lines. For example: *Arcturus*.
7. **M Star:** These are red stars with the surface temperature less than 3500 K. Their spectrum has strong CaI absorption lines. For example: *Betelgeuse*.

## 2.2 Morgan-Keenan Luminosity Classification Scheme

Harvard scheme cannot distinguish between stars of same temperature but different luminosities (brightness). Hence, there was a need of another classification scheme based on luminosities of stars. The M-K classification classifies the stars in the following categories:

1. **I:** Supergiants
2. **II:** Bright giants
3. **III:** Normal giants
4. **IV:** Subgiants
5. **V:** Main Sequence Dwarfs
6. **VI:** Subdwarfs
7. **D:** White Dwarfs

## 2.3 HR Diagram

The Hertzsprung–Russell diagram is a scatter plot of stars showing the relationship between the stars' absolute magnitudes or luminosities versus their stellar classifications or effective temperatures. We can get the information about the evolution and classification of stars from their HR Diagram. **This is the single most important diagram in the field of Stellar Astrophysics and Astronomy.**

### 3. Objectives

- To create an ontology based on Harvard Spectral Classification Scheme.
- To plot HR Diagram, Radius Histogram and Temperature Histogram to study the given stellar dataset.
- To create and test machine learning models on the stellar dataset for the following cases:
  - Two categories Morgan-Keenan Classification
  - Seven categories Harvard Spectral Classification

## 4. Methodology

### 4.1 Ontology

The term “ontology” comes from philosophy and corresponds to the “science of being”. This term was then used in computer science to designate a formal definition of all the objects in a domain and the relationships existing between these objects.

Ontologies are the structured knowledge which the computer can interpret and make judgements accordingly. To describe the knowledge, so that computer can work on it. Since we know that there is nothing obvious to a computer and we cannot make any prejudice on knowledge the computer have about the topic, we need to explain everything clearly.

Using Owlready2, We made an ontology that classify the stars in different spectral classes according to Harvard Spectral Classification Scheme on the basis of color, temperature and spectral lines.

Following is the description of the classes present in our ontology:

- **Color:** This class represents the color of the star. It has following sub-classes:
  - Red
  - Blue
  - Orange
  - Yellow
  - White
- **Temperature\_Range:** This class represents the temperature range of the star. It has following sub-classes:
  - Greater\_than\_25K
  - Between\_11K\_to\_25K
  - Between\_7K\_to\_11K
  - Between\_6K\_to\_7K
  - Between\_5K\_to\_6K
  - Between\_3K\_to\_5K
  - Less\_than\_3K

- **Dominant\_Spectrum\_Lines:** This class represents the dominant absorption lines in the star spectrum. It has following sub-classes:

HeII

HeI

HI

CaII

CaI

- **Stellar\_Spectral\_Class:** This class is the super class. The child classes of this class represent stars. Each of them has functional properties that defines the color, temperature and spectrum. It has following sub-classes:

O

B

A

F

G

K

M

We created individuals of this ontology. These individuals represents the actual stars present in the universe. These stars are: *Lacertra\_10*, *Rigel*, *Sirius*, *Canopus*, *Sun*, *Aructrus*, and *Betelquese*. Finally, we saved this ontology in the Results section.

## 4.2 HR Diagram

For the HR Diagram, we have taken stellar dataset available on [Kaggle](#). This dataset is a 6 class star dataset of 240 stars.

Following columns are present in the dataset:

- Temperature (K): Surface temperature of the star.
- Luminosity (L/Lo): Ratio of energy emitted per second by the star to the energy emitted per second by the sun.
- Radius (R/Ro): Ratio of radius of the star to the radius of the sun.
- Absolute Magnitude (Mv): Apparent magnitude of the star if it was viewed from a distance of 10 parsecs.
- Star type: This is a column that classify the stars in six different categories given below:
  - 0 : Red Dwarf
  - 1 : Brown Dwarf

- 2 : White Dwarf
- 3 : Main Sequence
- 4 : SuperGiants
- 5 : HyperGiants

- Star color: Color of the star
- Spectral Class: The spectral class of the star according to Harvard Spectral Classification Scheme The different classes are: O, B, A, F, G, K, M.

Before, we plot the necessary plots using matplotlib, we performed data cleaning on the given dataset. Some of the cleaning steps, we have taken are listed below:

1. The same type of color in `df['Star color']` has different names. Also, the color names were not following the syntax used by `matplotlib.pyplot`. To remove this, we made an array named `colors`, which will store the colors of all the stars in the proper syntax.
2. The radius of the stars have very low values (almost close to zero). These values cannot be used for sizes in `plt.scatter()`. The stars will not be visible on the plot. Hence, we binned the radius of the stars into different ranges. This discretized the size of markers on the HR plot.
3. We made the functions that correlate the Y-axis with the secondary Y-axis. This was required for accurate axis labels on the y-axis.

Finally, we have plotted the HR Diagram, Radius histograms and Temperature histograms. The inference we got from these diagrams were given in the Results section.

## 4.3 Machine Learning

The machine learning part of the project can be broadly classified into two sections:

1. **Dwarf-Giant Classification:** This includes the classification of the stars from the dataset into two categories i.e. Dwarf stars and the Giant stars based on the features available.
2. **Spectral Type Classification:** This includes the classification of the stars from the dataset into seven categories according to the Harvard Classification scheme.

To make any progress towards training the models and making a suitable and well functioning machine learning model, the data has to be cleaned and managed. Hence we performed the data cleaning via the following steps, chronologically.

### Creating the Data-frame

- We imported the CSV dataset named "Star99999\_raw" into a dataframe using the Pandas system in Python. As observed the dataset contained two columns containing the indexing of the stars, one intrinsic and the other generated by the Pandas. Since we needed just one indexing in the dataset, we dropped the excess column, thus reducing the size of the dataset by 1 column.



- Since the dataset contained values that can cause errors while trying to convert them into float, we add a parameter `errors='coerce'` to force the function to convert bad non-numeric values to NaN.

### Removing the Missing Data

- As we observed in the dataset, many of the rows contained missing values in some of the columns. This if left untreated might lead to bad machine learning and inefficient training leading to inaccurate results. Hence we traced out all the rows that had a missing value and calculated the percentage of such rows in the total dataset.
- The total percent of missing values was approximately 0.84% of the total values in the dataset. Since this was just a small fraction of the total, removing the rows did not have much effect on the total data available. Hence we dropped all the rows that had a missing values, which accounted in the total loss of 3% of the total rows in the dataset.
- Now since many rows from in between the dataset were removed, the indexing was disrupted. Hence we re-indexed the overall dataset and saved this new set into a new CSV file named "Data0".

### Absolute Magnitude Calculation

Absolute magnitude ( $M$ ) is a measure of the luminosity of a celestial object, on an inverse logarithmic astronomical magnitude scale. An object's absolute magnitude is defined to be equal to the apparent magnitude that the object would have if it were viewed from a distance of exactly 10 parsecs (32.6 light-years), without extinction (or dimming) of its light due to absorption by interstellar matter and cosmic dust. By hypothetically placing all objects at a standard reference distance from the observer, their luminosities can be directly compared among each other on a magnitude scale.

The absolute magnitude of the stars were generated via the equation:

$$M = m + 5(\log_{10} p + 1)$$

Where  $M$  represents the absolute magnitude  $Amag$ ,  $m$  represents the visual apparent magnitude  $Vmag$  and  $p$  represents stellar parallax  $Plx$ .

In this session, we will create a new column  $Amag$  to store  $M$ .

Things need to be aware:

- Taking log of 0 would result in a infinity, which is what we dont want to see. Hence, to fix this: Dropping rows with  $Plx = 0$ .
- Taking log of -ve numbers would result complex numbers, which is what we dont want to see too. Hence, to fix this: Taking Absolute value of  $Plx$ .

Hence we removed the rows that had the values of  $Plx$  as zero, and again re-indexed the dataframe. For the rows that have a negative value of the Parallax, we will use the absolute values for it.

The dataset after reduction and cleaning can be seen in the results section.

## Feature Engineering

If we examine the dataset, we can see that the scatter plots shows an interesting division of the space. Just a small number of stars would be mis-classified if we drew a simple curve in order to separate the two kind of stars. The idea is to expand our features by, for example, calculating the squared value of some columns and also performing some calculations between two columns in order to obtain a third one. Some of the features used by us were, square of all the values such as B-V index, AMag, and VMag. We also added there features to each other in various combinations to make new features.

## Models

Our dataset contained stars from three more types that do not belong to the Harvard classification scheme. Hence we removed the rows that contained any such stars. As the number of such instances was very less, we did not reduce the dataset to much extent.

After the feature engineering we split the dataset into test and training sets in the ration of 20 to 80. Hence the test set had 20,000 stars and the training set had 80,000 stars approximately.

We applied 3 models on the dataset to classify the stars, KNN, Linear Regression, and Support Vector Classification. The Confusion matrix for each of the model can be seen in the results section.

For the spectral stellar classification, we just applied the KNN classification and achieved a final accuracy of about 60% in classifying the stars into 7 spectral types.

## 5. Results and Conclusion

### 5.1 Ontology

We used Protege Software to get the metrics for the created ontology. The details of the metrics are as follows:

- **Metrics**

- Axiom 168

- Logical axiom count 106

- Declaration axioms count 62

- Class count 29

- Object property count 6

- Individual count 28

- **Class Axioms**

- SubClassOf 28

- EquivalentClasses 7

- DisjointClasses 4

- Hidden GCI Count 7

- **Object Property Axioms**

- InverseObjectProperties 3

- FunctionalObjectProperty 3

- ObjectPropertyDomain 6

- ObjectPropertyRange 6

- **Individual Axioms**

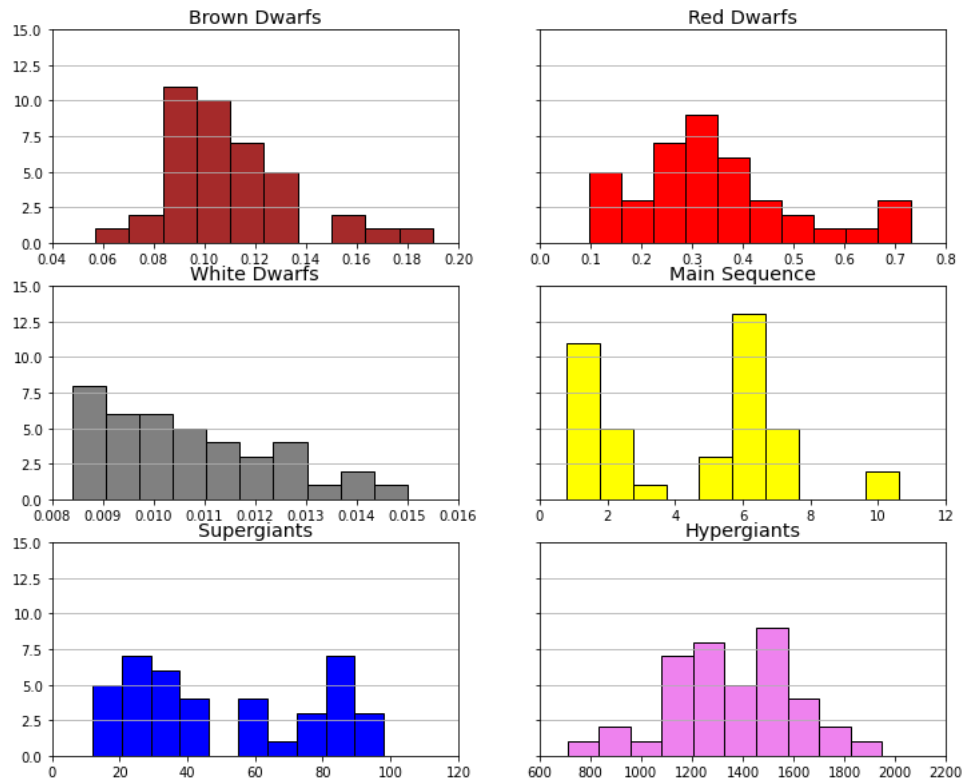
- ClassAssertion 28

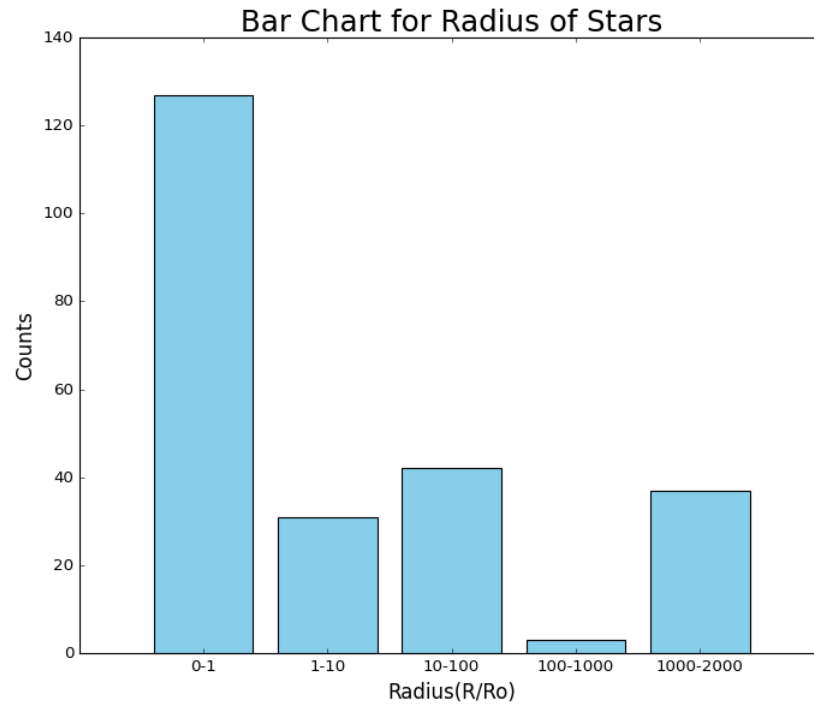
- ObjectPropertyAssertion 21

## 5.2 HR Diagram

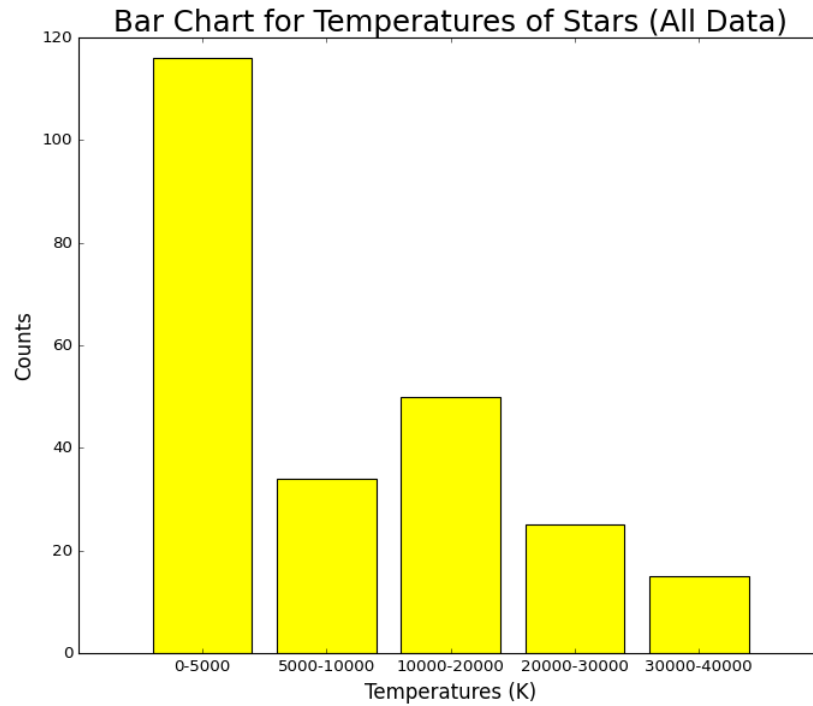
- We made the radius histograms according to star types, and an overall bar chart of radius of the stars. We found that the most number of stars have the radius in the range 0-1. This means that most of the stars in the dataset are smaller than the sun. This result coincides with the theoretical expectations. In the universe, most of the stars are dwarfs (smaller than sun).

Radius Histograms According to Star Types





- We plot the bar chart for the temperature of stars. We found that the most number of stars have the Temperature (K) in the range 0-5000. This means that most of the stars in the dataset have temperature less than the temperature of the sun (5800 K). We also infer that most of the stars belong to spectral class K and M.



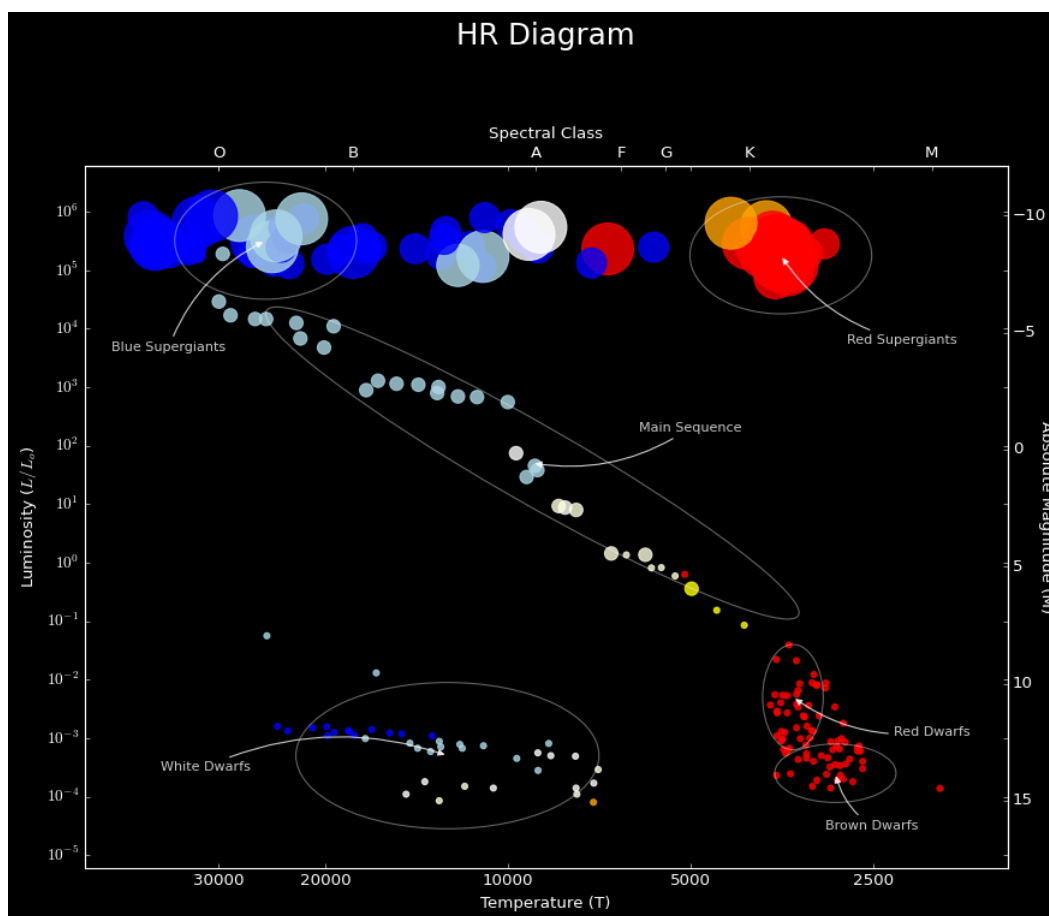
- We have plotted the HR Diagram for the given stellar dataset. All properties of the stars are represented in the above diagram.

The size of the markers is related to the actual radius of the stars.

The color of the stars is represented by the color of the markers.

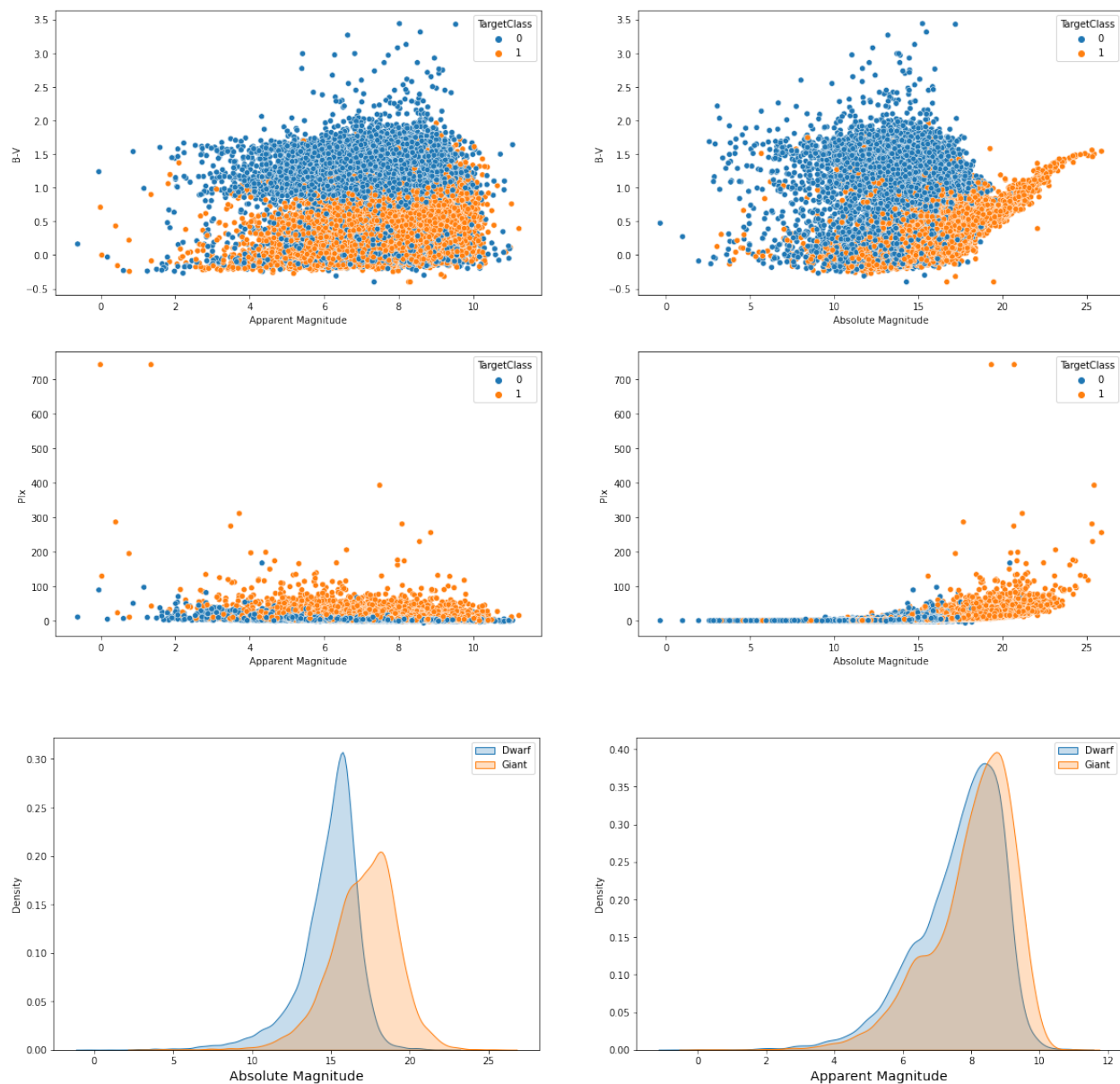
The X-Y axis give the information about the spectral class, temperature, luminosity and absolute magnitude of the stars.

Using the annotations, we labelled the regions containing the stars of different types.



## 5.3 Machine Learning

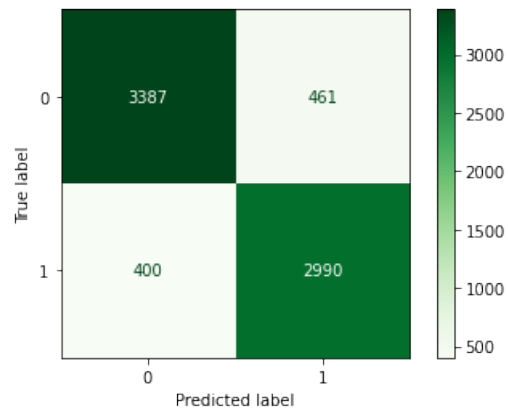
The dataset can be visualised below with different parameters from the dataset as the axis for the plots.



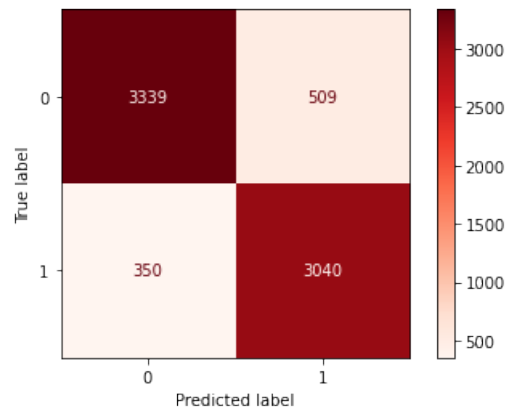
The confusion matrix for each of the models can be seen as:



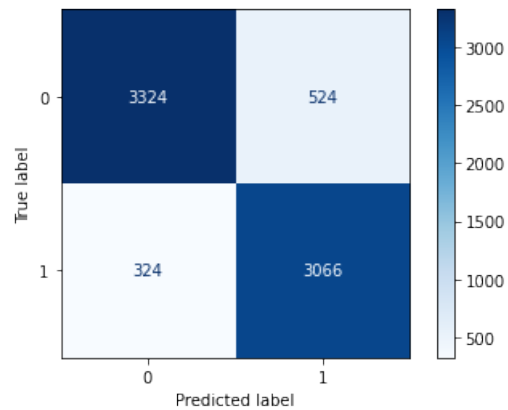
### 5.3.1 KNN Model: Dwarf-Giant Classification(Accuracy 88%)



### 5.3.2 Linear Regression Model: Dwarf-Giant Classification(Accuracy 89%)

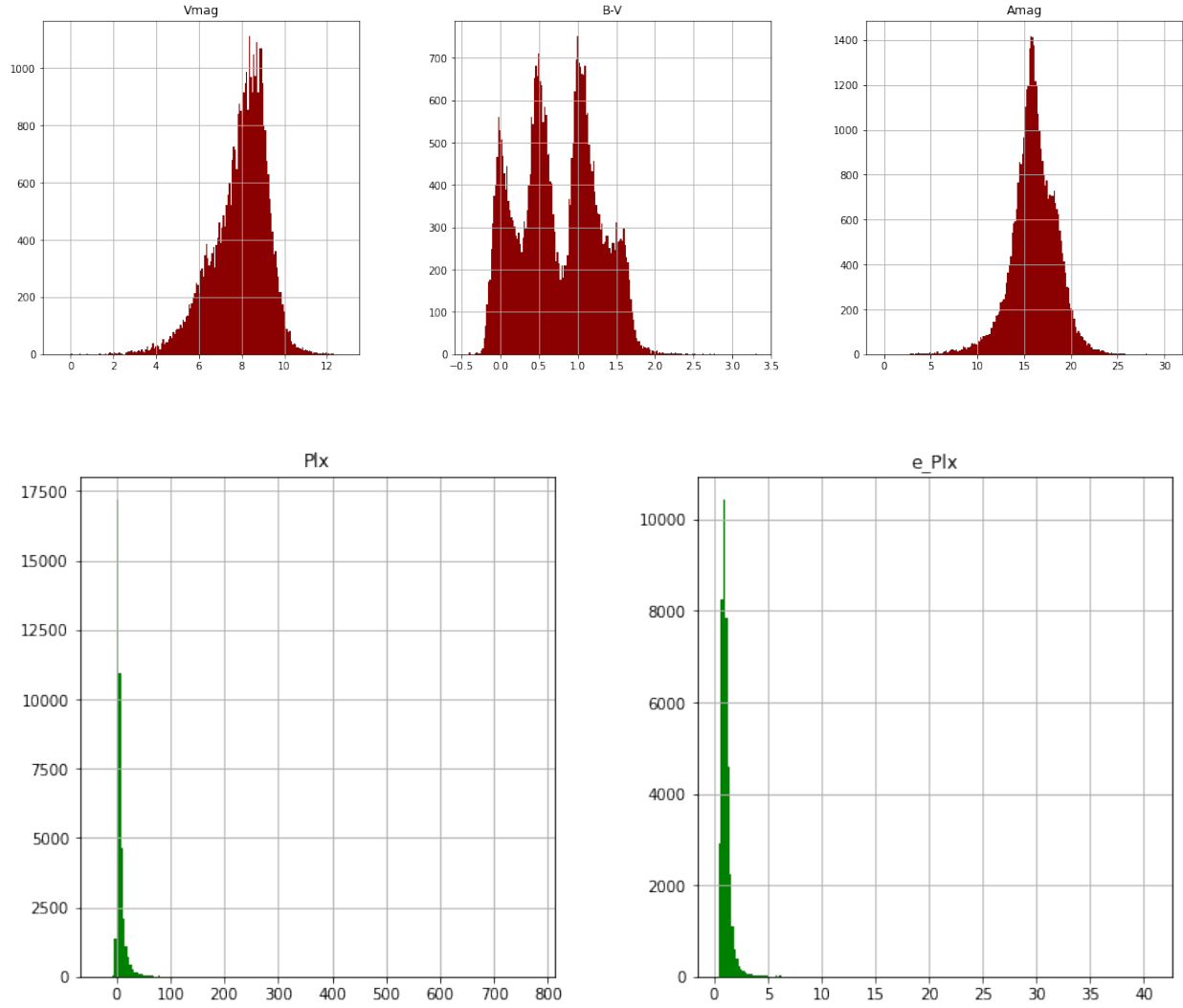


### 5.3.3 Support Vector Model: Dwarf-Giant Classification(Accuracy 88%)



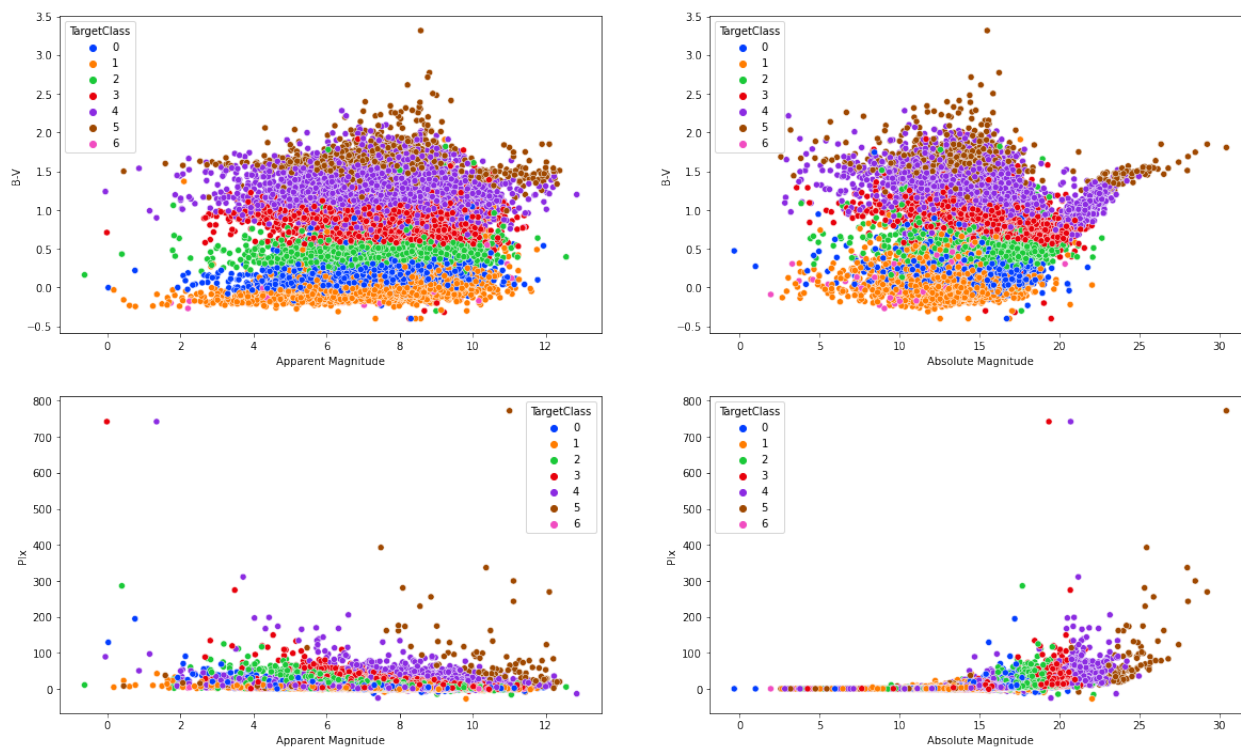
### 5.3.4 Data Visualization: Spectral Class

The dataset through the perspective of the features is seen as:

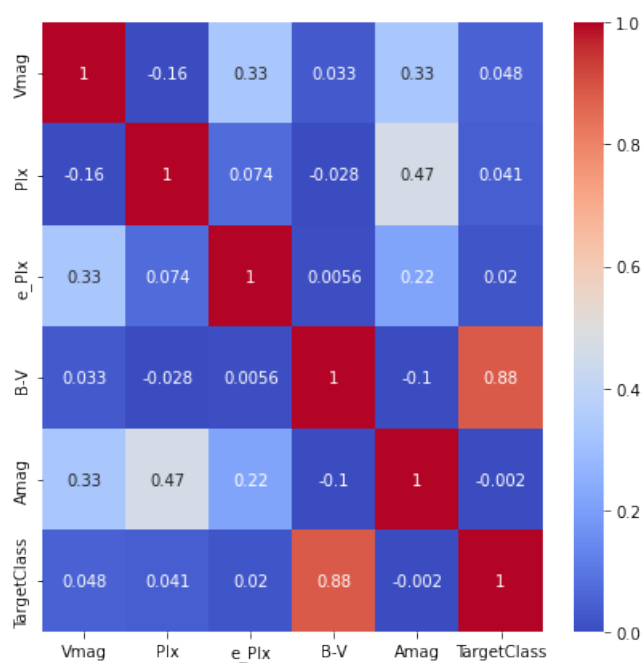


### 5.3.5 Spectral Types

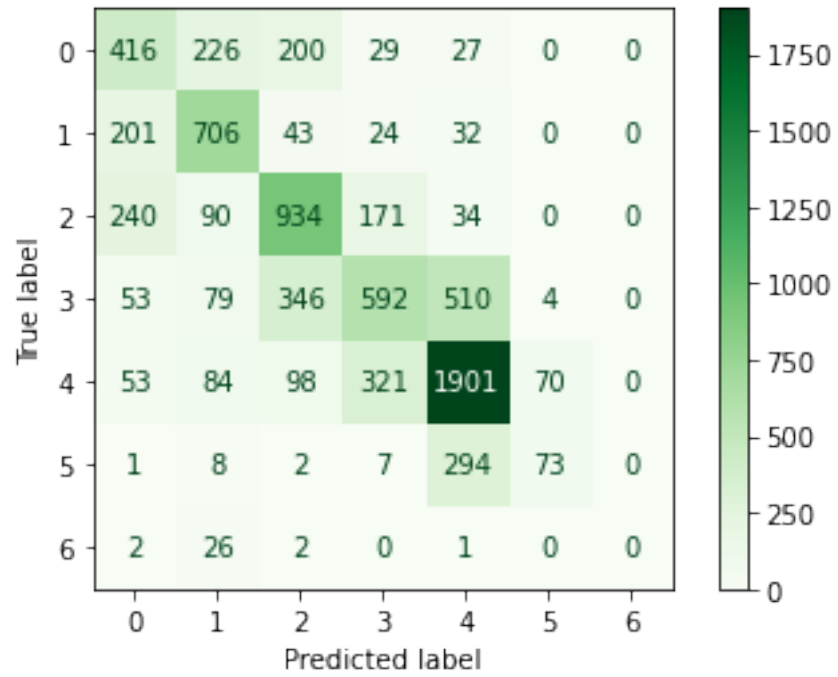
The overall dataset divided into seven spectral types can be seen as:



The heat map of the dataset is, which describes the overall distribution.



The confusion matrix for the last KNN model that classifies the data into 7 spectral types is:



## 6. References

1. [Classification of Stars from Redshifted Stellar Spectra utilizing Machine Learning](#)
2. [Stellar Dataset for Machine Learning](#)
3. [Stellar Dataset for HR Diagram](#)
4. [Harvard Spectral Classification Scheme](#)
5. [Morgan-Keenan Luminosity Classification Scheme](#)
6. [HR Diagram](#)