South Dakota Mines
Data Mining, Spring 2022

CSC 454

Stellar Data Mining - Starry Team

Stellar Data Mining First Deliverable Document

Mathew Clutter, Chami Senarath

1 Introduction

If one takes a look up at a dark night sky, one will find it to be illuminated by hundreds and thousands of twinkling stars. While they may look the same to untrained eyes, but a closer look at them reveals a diversity within the stars. Some may look redder or bluer than others; some may look bigger and smaller than others; some are young and some are old. This provided us humans methods to cluster as constellations and classify into spectral classes that helped to either navigate in the sea or navigate in space.

In astronomy, stellar classification is the classification of stars based on their spectral characteristics. Astronomers classify stars according to their physical characteristics. With either trained or clear enough skies, it is feasible to notice that stars display different types of colors. Since temperature and color are extremely closely related, it makes sense to classify by a study of color (photometry). These colors can be dissected into separate wavelengths and if there are either neutral or ionized atoms in the outermost layer of the star, they will absorb some of the light at particular wavelengths. These absorption information lead to the earliest classification system. This system is known as Secchi classes, named after the 19th century Italian Astronomer Angelo Secchi.

1. Class I: A class for the blue/white stars that exhibited strong, broad hydrogen lines
2. Class II: Yellow stars with weaker hydrogen features, but with evidence of rich metallic lines
3. Class III: Red stars with complex spectra, with huge sets of absorption features.

After several iterations of stellar classification revisions, currently Astronomers uses a classification system known as Morgan-Keenan (MK) system using the letters O, B, A, F, G, K, and M, a sequence from the hottest (O type) to the coolest (M type). Each letter class is them subdivided using a numeric digit with 0 being hottest and 9 being coolest. In this system, a luminosity class is added to the spectral class using Roman numerals. This is based on the width of certain absorption lines in the star's spectrum, which vary with density of the atmosphere.

The goal of this project is to understand if the current model of classifying stars able to determine the correct type when a new star is being discovered. This will be a valuable resource for Data Miners in the field of Astronomy. This analysis will be developed by Mathew Clutter and Chami Senarath. Their email addresses can be found below.

- Mathew Clutter: Mathew.Clutter@mines.sdsmt.edu
- Chami Senarath: Chamaka.Senarath@mines.sdsmt.edu

As both Mathew and Chami are outer-space enthusiasts, they are excited to work on a project related to stellar classification and expand their knowledge on such subjects.

2 Information about the Data Set

This data set was created used the following website: `https://vizier.u-strasbg.fr/viz-bin/VizieR-3?-source=V/137D/XHIP&-out.max=50&-out.form=HTML%20Table&-out.add=_r&-out.add=_RAJ,_DEJ&-sort=_r&-oc.form=sexa`

This website queries a large star database, and provides the attributes that are requested. For this classification project, the requested attributes consisted of the spectral type, temperature class, luminosity class, iron content, age, number of exoplanets, B-V magnitude, absolute magnitude, and luminosity. This website allowed us to obtain a large data set containing the attributes necessary to classify stars into their appropriate spectral types.

This data set contains 117955 instances. There are 11 features, and 1 meta attribute. The 11 features and meta attribute are as follows:

1. Name: The common name of the star, if known (meta attribute)
2. Right Ascension: The right ascension of the star in the sky (component of the position of the star)
3. Declination: Declination of the star in the sky (another component of the position of the star in the sky)
4. Spectral Type: The spectral type (Morgan-Keenan System) that the star is a member of. This is the target attribute that we will be attempting to classify.
5. Temperature Class Codified: The temperature of the star, with values scaled to a range 0-147, with a median of 50, and mean of 48.9. The temperature range will allow comparisons of temperatures between stars.

6. Luminosity Class Codified: The luminosity of the star, with values scaled to a range 0-126, with median 4, and mean 4.1. Luminosity class allows for comparisons of luminosity between stars.
7. Iron Abundance: The amount of iron present in the star, if known. Compared to the amount of iron that is present in the sun.
8. Age: The age of the star, in billions of years.
9. Number of Exoplanets: The number of exoplanets that the star has, if known.
10. B-V Color Index: Describes numerically the color of the star, according to the Johnson Color Index.
11. Absolute Magnitude: The absolute magnitude of the star.
12. Stellar Luminosity: The luminosity of the star, compared to the luminosity of the sun.

The following shows an overview of the summary statistics for each of the attributes:

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
_RAJ2000 deg		181.466	181.97	0.569038	0.000900716	359.979	0 (0%)
_DEJ2000 deg		-2.14329	-1.94596	-19.137	-89.7824	89.5694	0 (0%)
Tc		48.8744	50	0.279963	0.00	143	3717 (3%)
Lc		4.12009	4	1.01363	1	126	44033 (37%)
[Fe/H] sun		-0.0880458	-0.06	-10.9905	-4.8	77	98812 (83%)
age Gyr		3.79028	2.5	0.875049	-2.48	72	105009 (89%)
Npl		1.18624	1	0.665248	0.00	6	117519 (99%)
B-V mag		0.705365	0.612	0.692173	-0.4	5.46	1290 (1%)
Kmag mag		1.04948	1.49	2.13618	-10.26	13.35	58984 (50%)
Lum Lsun		91.9669	9.15	12.6836	-7.45	125579	58208 (49%)

Figure 1: Summary Statistics

As this is a large data set, and not every star has every property recorded, there are some missing values. However, the most important attributes for classification are mostly present.

Histograms and box plots for each numerical attribute follow:

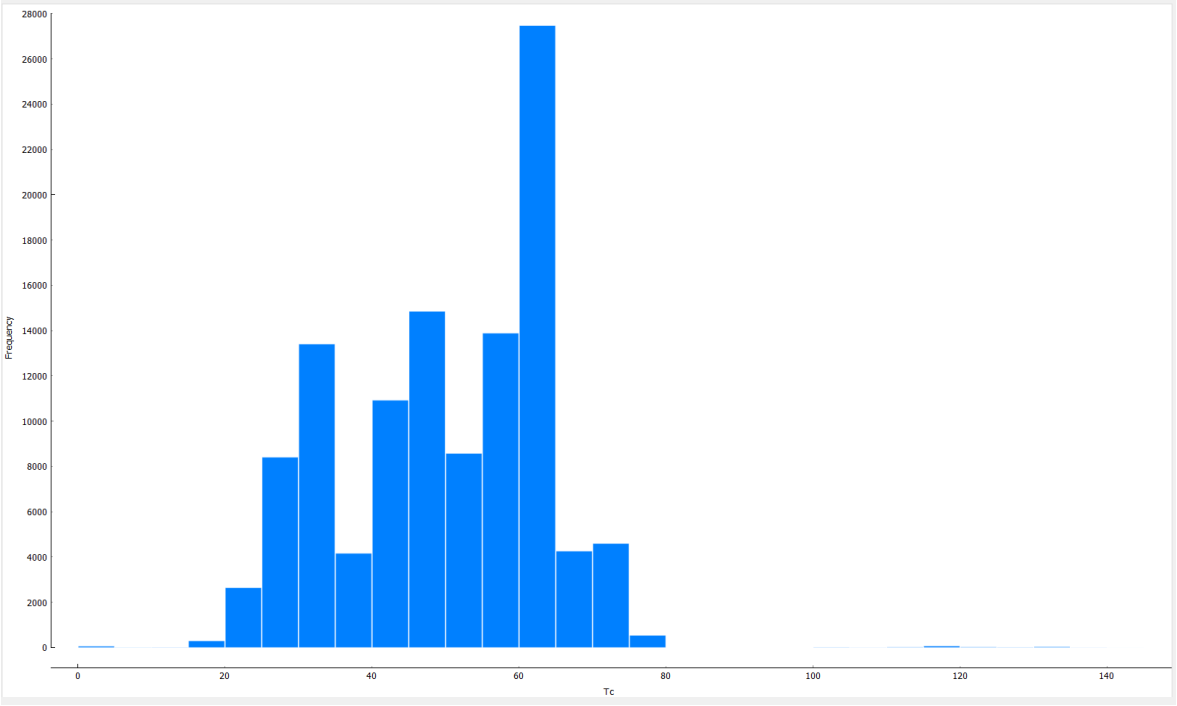

Figure 2: Tc Histogram

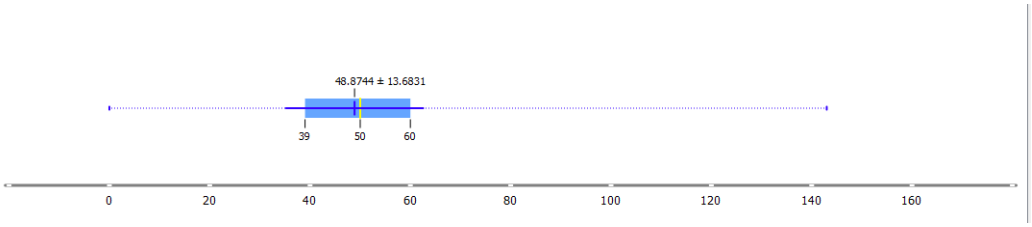

Figure 3: Tc Box Plot

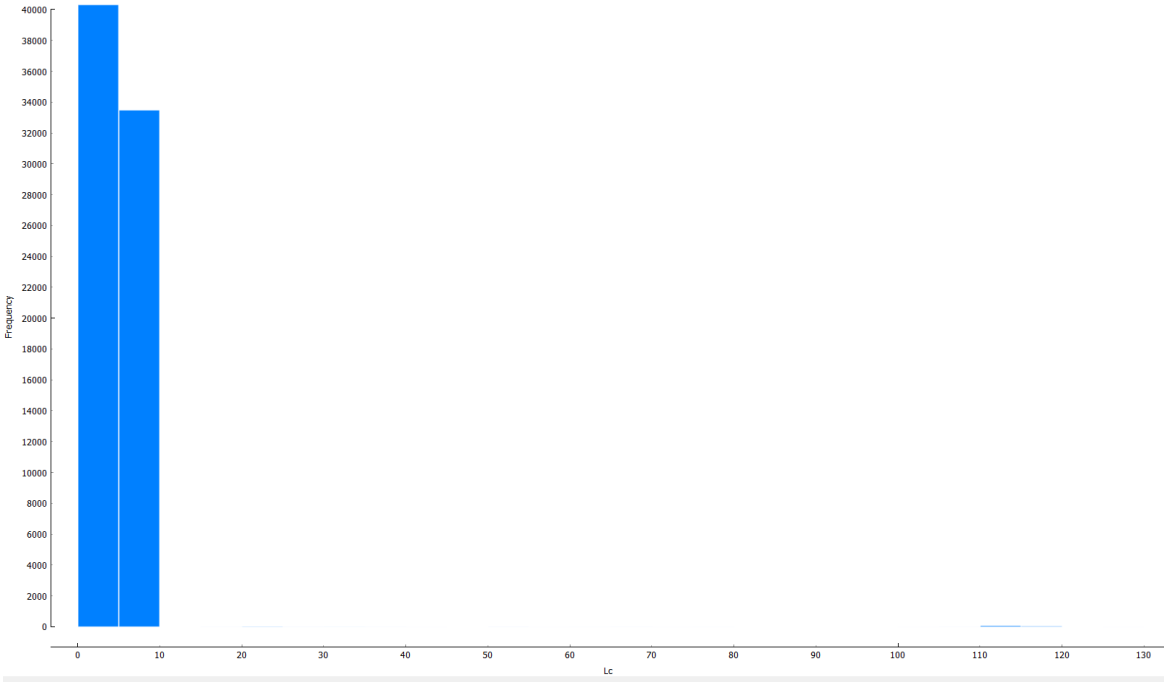

Figure 4: Lc Histogram

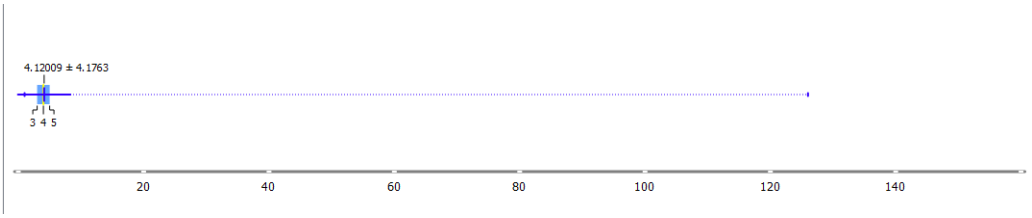

Figure 5: Lc Box Plot

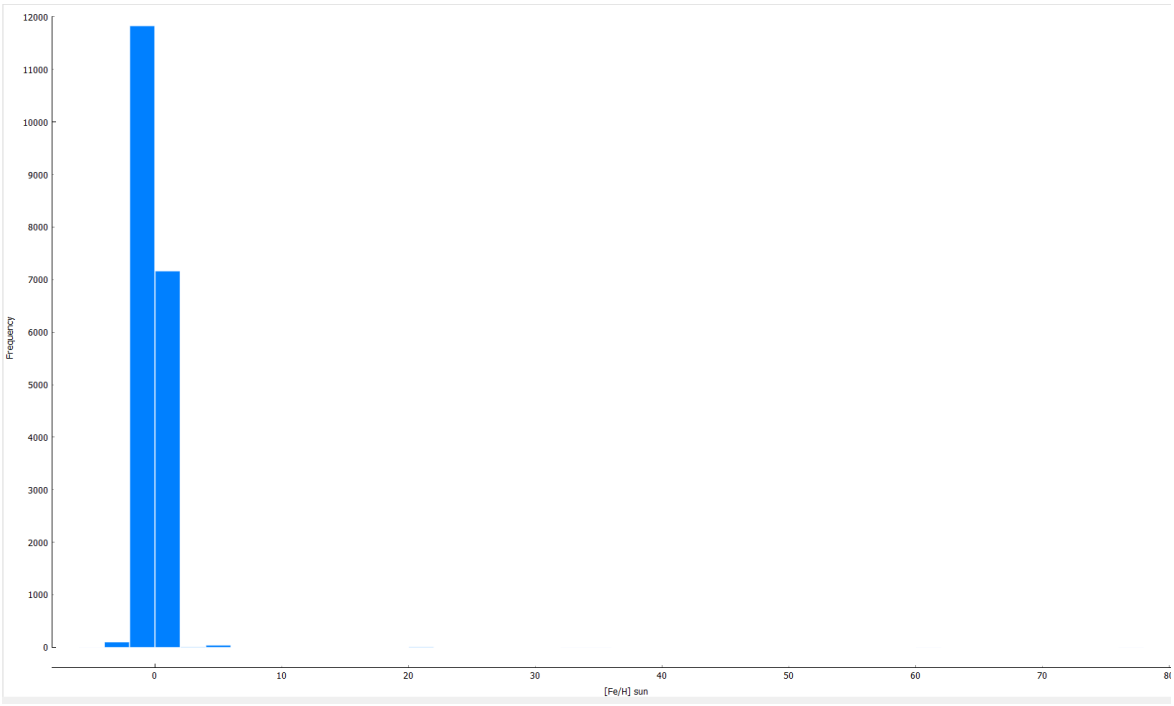

Figure 6: Iron Content Histogram

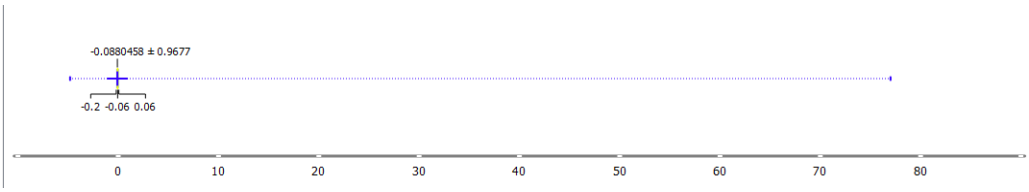

Figure 7: Iron Content Box Plot

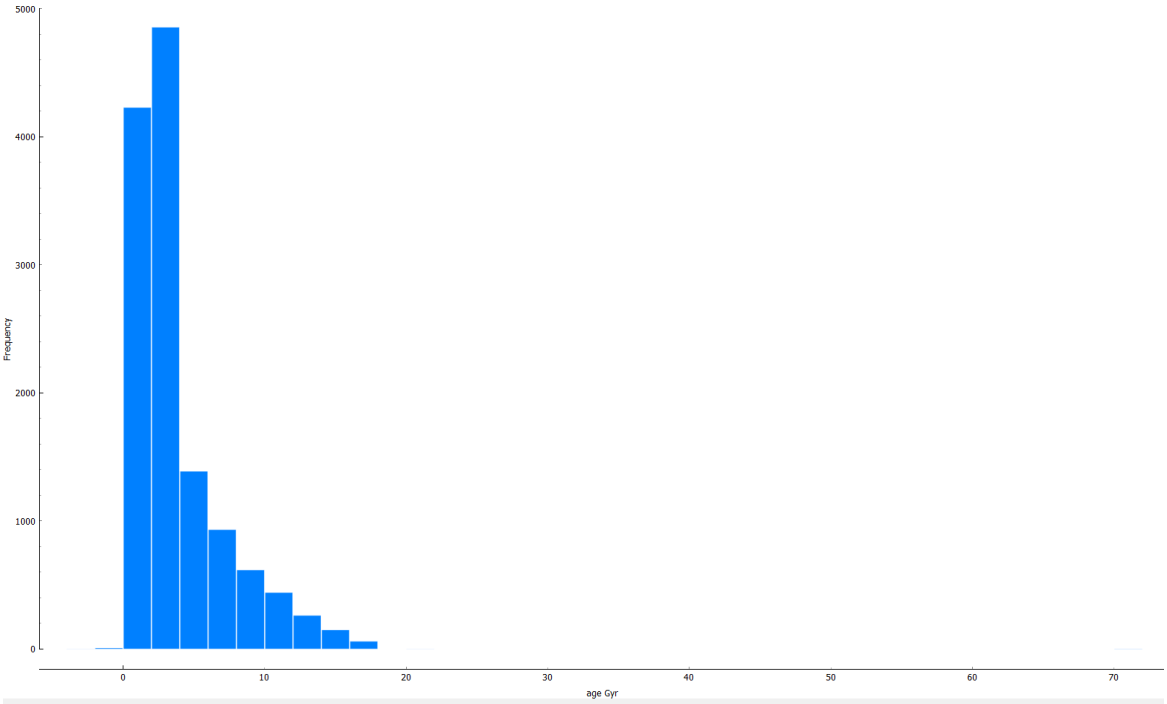

Figure 8: Age Histogram

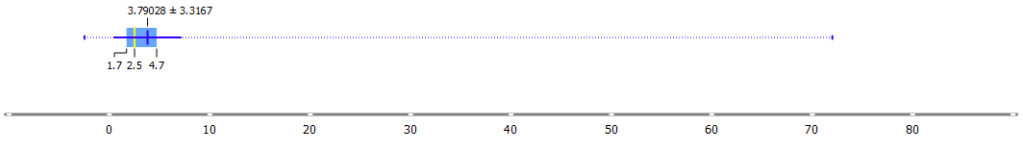

Figure 9: Age Box Plot

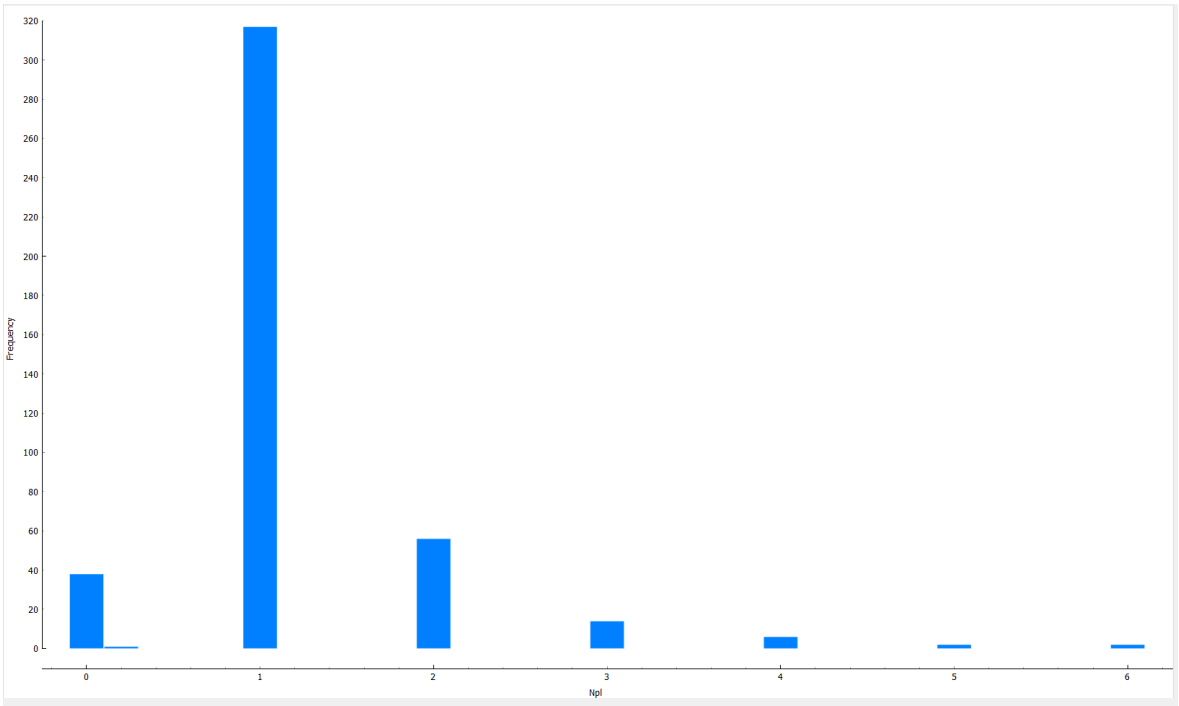

Figure 10: Exoplanets Histogram

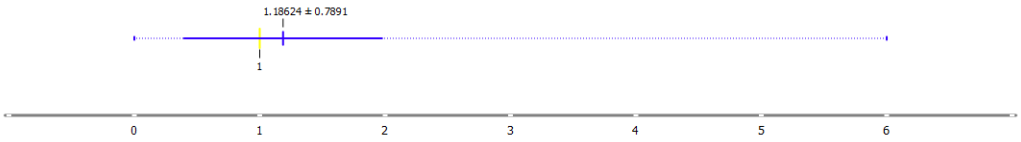

Figure 11: Exoplanets Box Plot

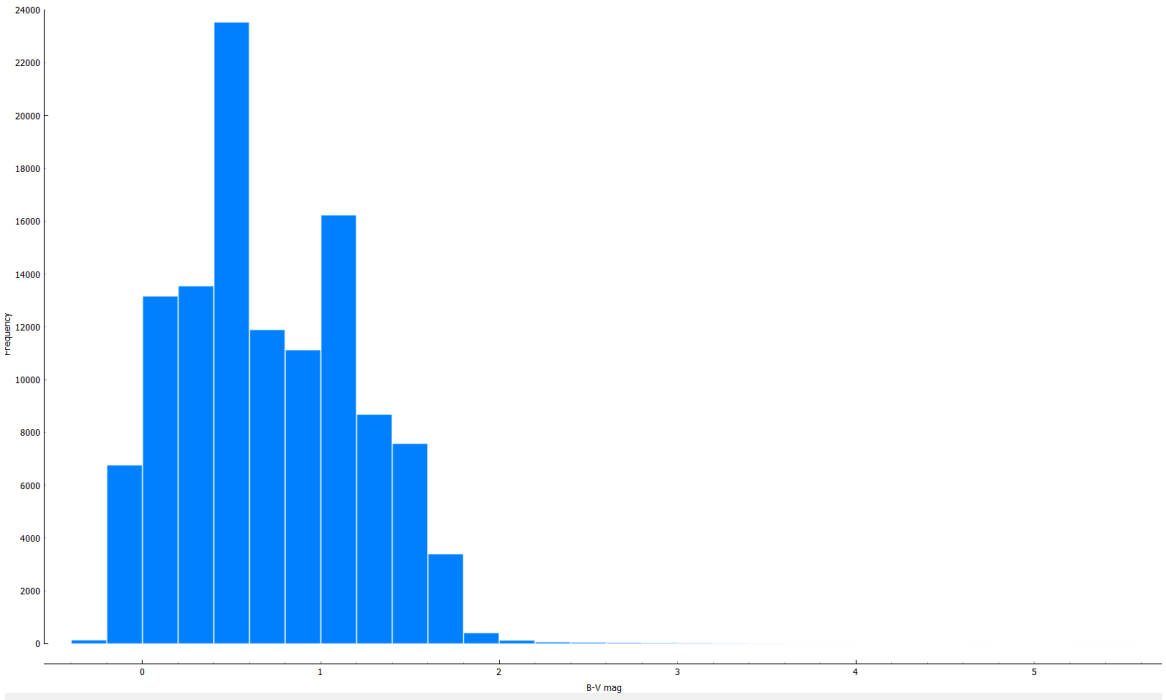

Figure 12: B-V Color Index Histogram

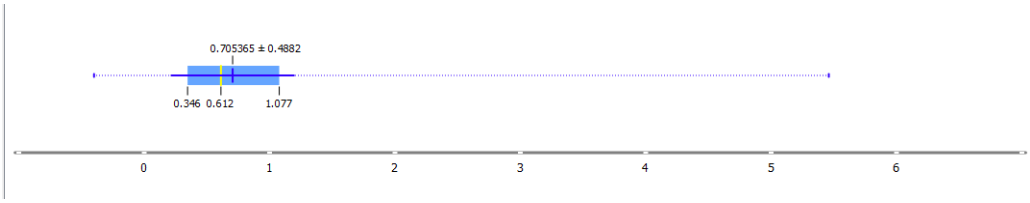

Figure 13: B-V Color Index Box Plot

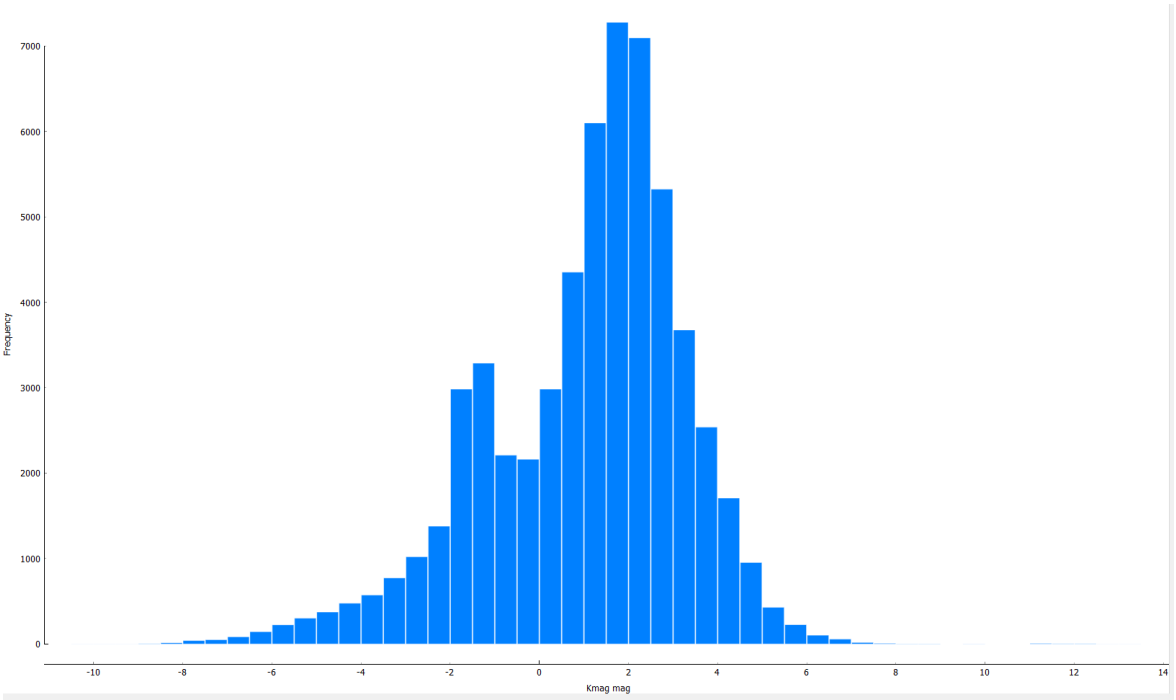

Figure 14: Absolute Magnitude Histogram

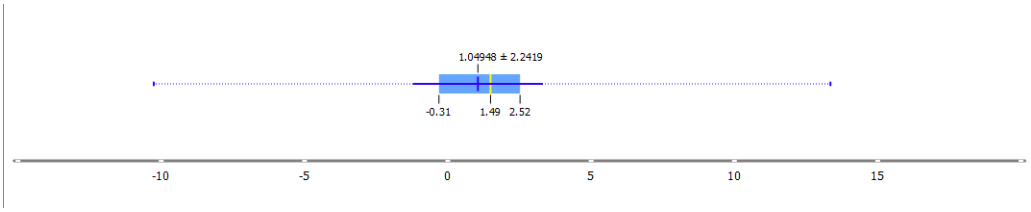

Figure 15: Absolute Magnitude Box Plot

Figure 16: Stellar Luminosity Histogram

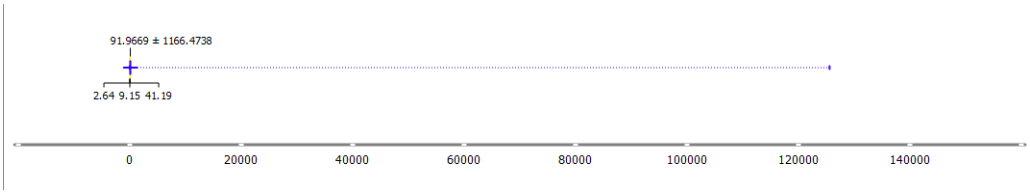

Figure 17: Stellar Luminosity Box Plot

As we can see, the raw data needs a lot of preprocessing to make some sense. Hence, we will walk through each attribute to discard unnecessary data and remove outliers. Also, we may impute to fill in some of the missing values.

3 Ethical and Privacy Issues

Since the data set we are using is related to stars, there are no ethical issues within the data that would be a concern. The data set contains no personal information, hence there are no privacy issues that will be mingled.

4 Our Hypothesis

In class, we learned about 4 types of classifiers. They are as follows:

1. Decision Tree Classifier
2. Rules-Based Classifier
3. Nearest-Neighbor Classifier
4. Naïve Bayes Classifier

Our hypothesis is that a Decision Tree Classifier would be the most effective to determine a classification class type for a star. The reason for this hypothesis is as follows: There are several missing values in several attributes that can bring out poor performance in Rule-based and Naïve Bayes classifiers. Some attributes can be considered as irrelevant, which might influence some classifiers to perform poorly. However, the Decision Tree Classifier can handle such features. A downside of it would be that it performs poorly when there are attributes that interact with each other. For examples, temperature, color index, and stellar luminosity are closely related but other classifiers posses that disadvantage too.

5 How the Study will be Conducted

We will use Orange to classify stars according to the above mentioned 4 different types of classifiers. Since we have 117955 instances within our data set, we have to select a certain percentage of instances for training data and test data. The percentage split will be determined in a such way that we are not overfitting. We will have to remove some outliers in our data set to make sure that incorrect rules will not be generated. Along with this, we will determine if there are irrelevant attributes. Then for each classification system, we will calculate performance metric (accuracy) and error rate to determine which classifier performs the best. We will experiment with and analyze the performance of the different classifiers using different evaluation methods, such as cross validation, bootstrap, and stratification. Through this analysis, we will attempt to determine the best classifier for this data set, to predict the spectral type.