# Star Types Dataset Visualization and Classification Using Machine Learning

Ahmed Bera Pay
*Department of Computer Engineering*
*Boğaziçi University*
Istanbul, Türkiye

***Abstract*—As fundamental celestial bodies, stars have intrigued astronomers for centuries towards new dimensions in understanding stellar evolution and cosmic dynamics, which also serve as cornerstones in understanding the universe's evolution. In this context, various techniques and methods have been used to classify stars based on features such as temperature, luminosity, and magnitude. In the contemporary era, machine learning emerges as a powerful and comprehensive concept, used in astronomy to process astronomical datasets and train models for predictive analysis. This paper investigates the relationships, correlations, and significance of different parameters, including luminosity, temperature, and spectral characteristics, in classifying different star types. This research employs some of the machine learning algorithms, namely Random Forest, kNN, SVC, XGB, and Logistic Regression, to construct prediction models to classify stars such as Red Dwarf, Main Sequence, Super Giants, etc. The Random Forest algorithm exhibited the highest accuracy in correctly predicting the star types.**

***Index Terms*—machine learning, stellar classification, stars, Hertzsprung-Russel Diagram**

## I. Introduction

The observation and classification of stars have been of paramount importance in astronomy, for centuries as a result of human curiosity about the cosmos. Traditionally, astronomers have utilized a variety of techniques to observe and classify stars, based on characteristics such as temperature, luminosity, and spectral features to classify them into distinct types. This manual classification, while yielding valuable insights, has some challenges in handling the big and complex datasets generated by modern astronomical observations.

The utilization of machine learning algorithms has revolutionized the field of star classification as many other fields. The capabilities of machine learning models to learn about the patterns and relationships within an extensive dataset offer a good framework to enhance the accuracy and efficiency of star classification. The utilization of these techniques is not only motivated by the volume of the astronomical dataset but also to seek an opportunity to explore some useful insights that might be difficult to detect with traditional methods.

This study will present a basic example of star classification, employing some machine learning techniques, including Random Forest, kNN, SVC, Logistic Regression, and XGBoost. The study also involves many visualizations of the star dataset using various techniques to examine the features and relationships within the data. This relatively small yet comparative analysis aims to provide a fundamental framework to better understand the characteristics of stars using basic data visualization and machine learning techniques which can be extended for many different celestial bodies using much more advanced machine learning techniques.

### A. Harvard Star Classification Scheme

Stars are categorized based on their spectral characteristics, with electromagnetic radiation from each star analyzed on a spectrum. Different chemical compositions result in different spectral lines that can be observed. Historically, stars were classified using the one-dimensional Harvard system, which relies only on the surface temperatures of stars. Various letters are assigned to describe different temperature ranges.

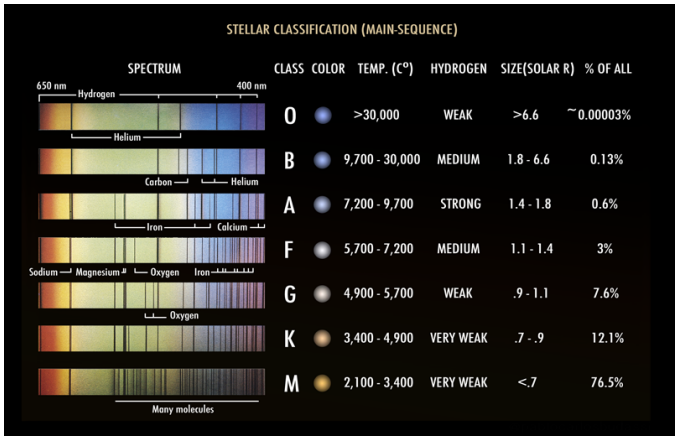


Fig. 1. Stellar classes based on Harvard scheme [1]

Additionally, each letter is further divided into numerical values to provide more detailed temperature descriptions. The modern Morgan-Keenan classification system incorporates luminosity measurements alongside temperatures for a more comprehensive classification. Luminosity classes are expressed using Roman numerals. [1]

In this study, the focus will be on classifying stars as main sequences, dwarfs, and giants, rather than using their spectral classes. Relationships with the spectral class will be visualized later in the study.

### B. Hertzsprung-Russell Diagram

The Hertzsprung-Russel diagram serves as a scatter plot of stars, illustrating the relationship between their absolute magnitudes and luminosities in comparison to their stellar classifications and temperatures. Fig. 2 and Fig. 3 represent different types of stars and their characteristics, helping to a better understanding of this study's concept and the evolution of stars. [3]

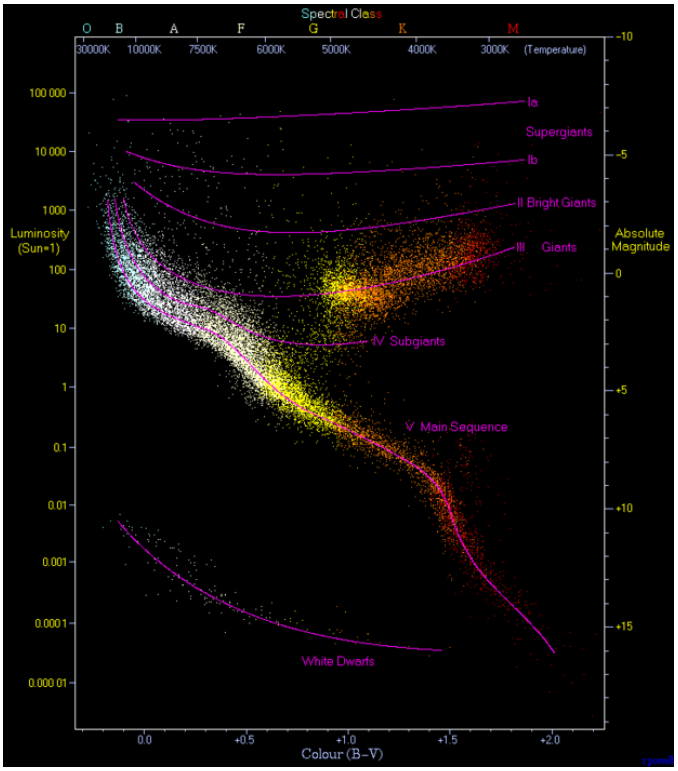


Fig. 2. [3]

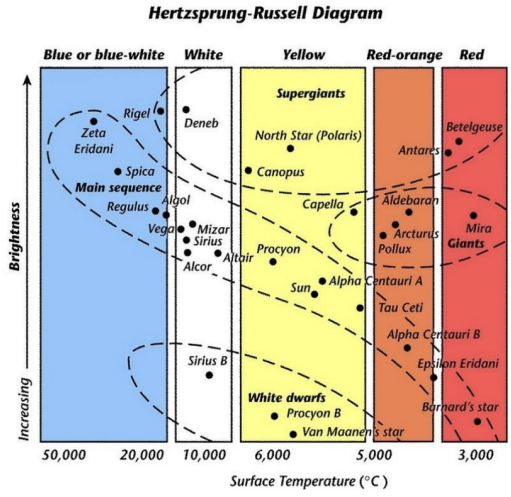


Fig. 3. [2]

The star types that can be seen in Fig. 2 and Fig. 3 will be predicted using machine learning models in this work.

## II. Methodology

### A. Data

The dataset subject to the study is found from Kaggle. The data were mainly collected from the web and the missing data were calculated using the following.

- Stefan-Boltzmann's law of Black body radiation (To find the luminosity of a star)
- Wienn's Displacement law (for finding the surface temperature of a star using wavelength)
- Absolute magnitude relation
- Radius of a star using parallax. [4]

Therefore, the dataset was complete and mainly ready to work on. The dataset is relatively small containing only about 200. This situation causes overfitting while facilitating the comparisons and the examinations. However, the dataset is still useful since the primary objective of this study was to analyze the characteristic data of the stars and train some models for them. Larger datasets can be used to build much better prediction models. The dataset contains many attributes regarding the stars. A brief explanation for some of them is as follows:

- **Temperature (K):** The temperature attribute describes the surface temperature of the star in Kelvin
- **Luminosity (L/Lo):** The luminosity attribute describes the star's luminosity (L) relative to the Sun's luminosity (Lo) which is roughly $3.846 \times 10^{26}$ Watts.
- **Radius (R/Ro):** The radius attribute describes the star's radius (R) with respect to the Sun's radius (Ro) which is roughly equal to 6.9551 x $10^8$ meters
- **Absolute Magnitude (Mv):** Absolute magnitude is another measure of luminosity for stars. In order to facilitate the use of measurement values in calculations, a logarithmic quantity is used and absolute magnitude is the apparent magnitude a star would have if it were seen from a distance of 10 parsecs. [12]
- **Spectral Class:** The spectral class attribute describes which spectral class a star belongs to according to the Harvard Stellar Classification Scheme.
- **Star Type:** The star type attribute describes the category in which the star takes place. This attribute will be the target parameter to compare the stellar features and build models. The models used in this work try to predict the type of a star given its aforementioned attributes.

### B. Models

This work employs various machine learning models to predict the star type accurately and also compares their results. The models utilized are as follows:

*1) Random Forest Classifier:* Random Forest is a widespread machine learning algorithm that combines the results of different decision trees to have a single output. Each tree is constructed on a bootstrapped sample of data, and a

random subset of features is considered at each split. In that way, the predictive performance is increased alongside reducing overfitting. The final prediction is made by aggregating the predictions of these individual trees. [5] [6]

*2) k-Nearest Neighbor Classifier:* The k-Nearest Neighbors is a simple, yet effective supervised machine learning algorithm used for classification. In this approach, each data point is characterized by its features. When a new data point is presented, the algorithm identifies the k-nearest data points in the space. The class occurring most frequently in the neighborhood is set as the class of the presented dataset. The distance between data points can be calculated using Euclidean distance. The algorithm offers a simple technique, but its performance is influenced by the choice of the "k" parameter and the feature scaling. One way among many ways to choose the k value, is running the algorithm many times with different k values to choose the one with the best outcomes. [7] [8]

*3) Support Vector Classifier:* The support Vector Classification algorithm aims to find an optimal hyperplane to separate the data points of different classes in high-dimensional space. The hyperplane is positioned to maximize the margin between the hyperplane and the nearest data points of each class. These are known as support vectors. This algorithm has the capability of capturing non-linear relationships which makes it a good candidate for the datasets in which the relationships are complex. [9]

*4) XGBoost Classifier:* XGBoost is a powerful gradient-boosting algorithm that sequentially builds decision trees, each correcting the errors of the previous one. This technique offers more speed and scalability than the traditional decision tree algorithm. [10]

*5) Logistic Regression:* The logistic regression algorithm applies the logistic function to a linear combination of input features. Instances are classified according to the output probability's fall in thresholds. The algorithm includes simplicity while having a limited capacity for capturing complex relationships. [11]

### C. Language

Python has been chosen as the programming language for this work due to its ease of use and extensive support from various visualization and machine learning libraries. The development process is done within a Jupyter notebook, providing a step-by-step demonstration along with relevant explanations.

## III. Results

The results of this work can be categorized into two sections: the examination/visualization of the dataset and the training with various machine learning algorithms.

### A. Data Visualizations

*1) Encoding and Histogram:* It starts with a basic analysis of the dataset with a univariate analysis which demonstrates the count of each star type in the dataset as a histogram in Fig. 5. For the rest of the paper, the star types will be addressed with their encoded versions. Here is the table for encodings:

| Encoding | Description |
|---|---|
| 0 | Red Dwarf |
| 1 | Brown Dwarf |
| 2 | White Dwarf |
| 3 | Main Sequence |
| 4 | SuperGiants |
| 5 | HyperGiants |

Fig. 4. Star type encodings

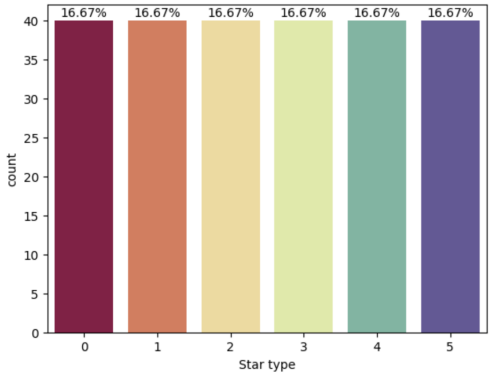


Fig. 5. Histogram for star types

The dataset has a perfect balance between all the star types. Having an imbalanced dataset could lead to bias from a particular star type. The equal representation does not require additional balancing techniques.

*2) KDE Plots:* The general density distributions of the features across the whole dataset are shown in Fig. 6 From the figure the most common values for each feature can be seen. The plots give insights into the value ranges for temperature, luminosity, radius, and absolute magnitude.

*3) Normality Test:* It is important to see whether the data follows a normal distribution before making inferences. In order to detect potential outliers and better understand the shape of the data distribution a normality test is conducted and resulted as in Fig. 7

The dataset mostly follows the normal distribution. Data for some stars like Red Dwarf and Hyper Giant seem a little off the normal distribution. But generally, the dataset is not far away from the normal distribution. Since the dataset size is quite limited, different results may occur when more data is available.

*4) Distribution Plots:* Plotting a distribution plot for each star type versus the features of temperature, luminosity, radius, and absolute magnitude as in Fig. 8 gives some idea regarding their relationships. The following can be deduced from these plots:

- The temperature appears to be a distinctive feature among

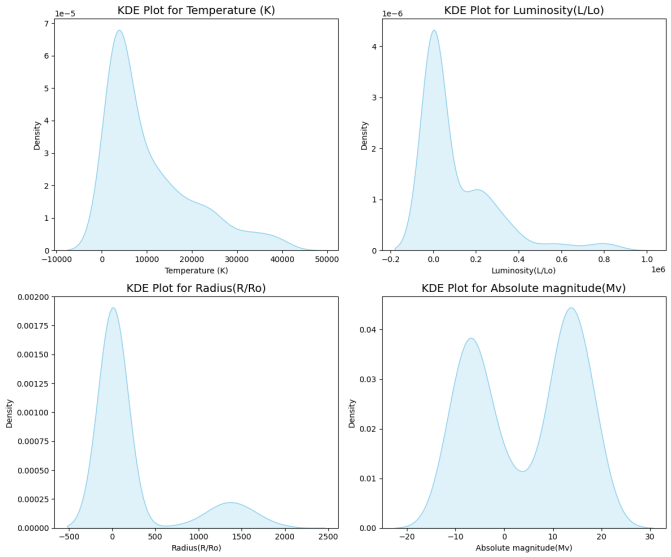


Fig. 6. Kernel density estimation plots for temperature, luminosity, radius, absolute magnitude

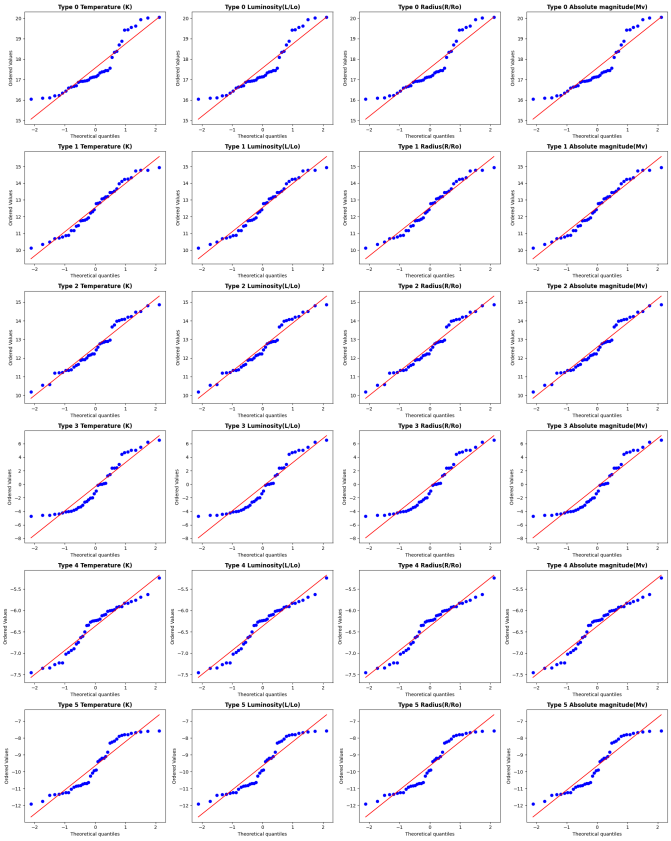

Fig. 7. Normality test for star types versus temperature, luminosity, radius, and absolute magnitude features

star types. Red dwarfs have the lowest temperature, around 3200K, while Giants exhibit the highest temperatures, surpassing 30000K.

- Similarly, luminosities also characterize stars, with a significant multiplicity difference between dwarfs and giants. Absolute magnitude, as another luminosity measurement, generally follows the same trend. The differing appearance of curves might be attributed to the logarithmic nature of absolute magnitudes.
- The radii of stars seem to be closely related to their types and other features such as temperature and luminosity. These outcomes align with the intuitive distinction between dwarfs and giants, as suggested by their names.
- Considering only these plots, it can be deduced that temperature, luminosity, and radius are positively correlated. Further analysis will help to better understand.

*5) Bivariate Analysis:* The comments above are supported by Fig. 9 describing the bivariate analysis and correlations between the star types and their features.

*6) Scatter Plots:* The relationship between the star colors or spectral classes and temperature, luminosity, radius, and absolute magnitude can be examined with scatter plots in Fig. 10.

The first row of scatter plots illustrates the temperature distribution concerning star color and spectral class. Some outliers and non-homogeneous distributions seem to exist, nonetheless, some conclusions can be drawn.

It can be observed that the highest temperatures are measured for blue stars, decreasing towards the colors of higher wavelengths. Yellow and red colors exhibit the lowest temperatures, aligning closely with the Harvard scheme in Fig. 1.

Similarly, the spectral class distribution concerning temperatures also aligns with the Harvard scheme. The highest temperature values are associated with spectral class O, while the lowest measurements are attributed to K and M spectral classes, as shown in Fig. 1.

For the luminosity feature, the distinctions are not very clear. The clearest results are obtained for red and blue colors, following the same trend as the temperature. Similarly, the spectral classes also follow the same trend as the temperature. Here, it can be concluded that temperature and luminosity are positively correlated, but relying solely on this limited dataset may lead to misconceptions.

The distributions for the radii of stars seem inconsistent with Fig. 1. While the maximum and minimum distances align with the expected colors and spectral classes, the points are widely spread. Based on these data, it appears challenging to draw a conclusion about the correlation between star colors and their radii. However, a better correlation between star types and radii is obtained in sections III-A9 and III-A10.

Since the absolute magnitude feature is another representation of luminosity, similar trends can be observed in the plots. Notably, the distributions here are clearer and more precise. Especially the relationship between spectral classes and absolute magnitude mostly aligns with Fig. 1. Apparently,

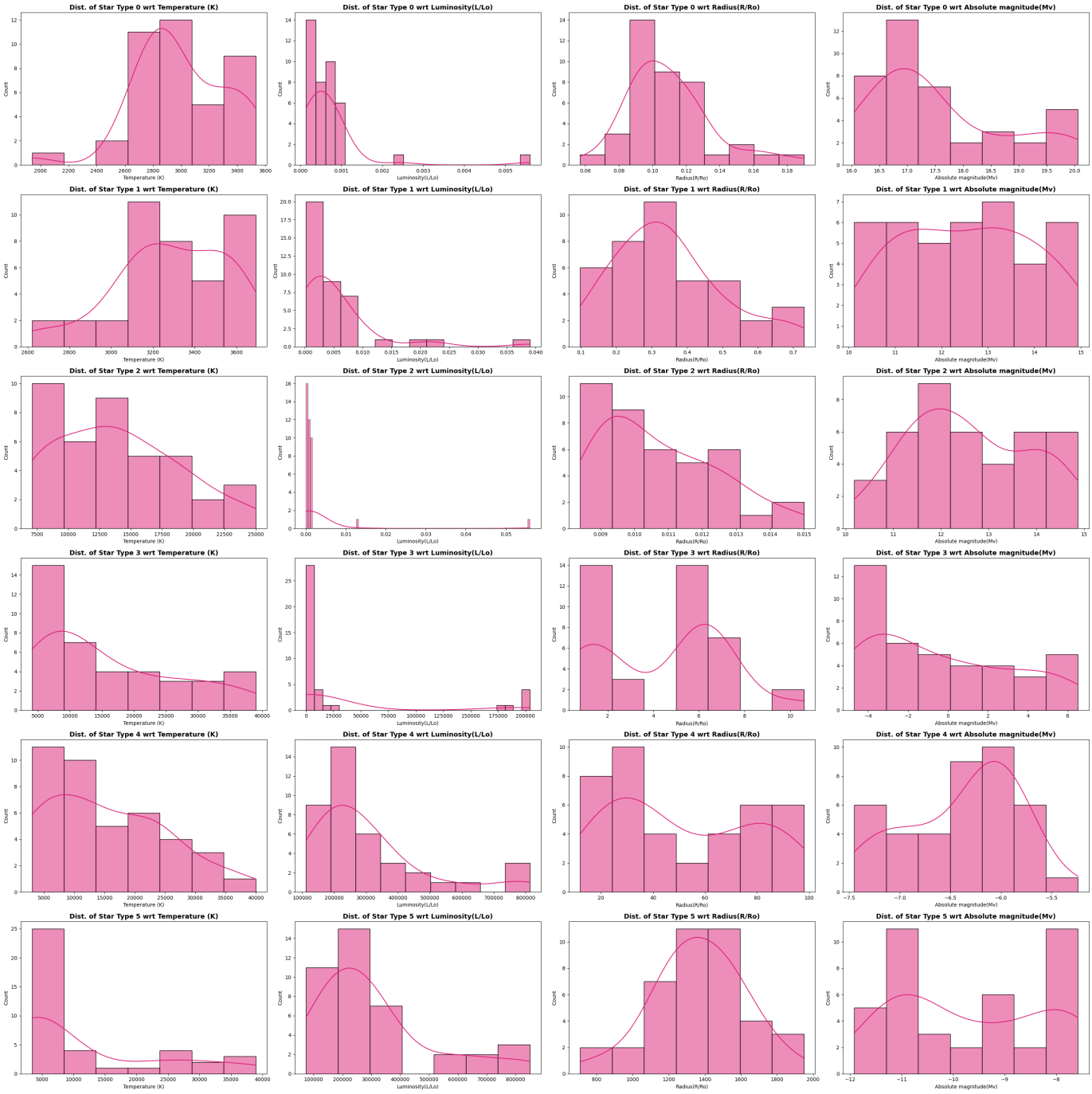


Fig. 8. Distribution plots of start types versus temperature, luminosity, radius, and absolute magnitude

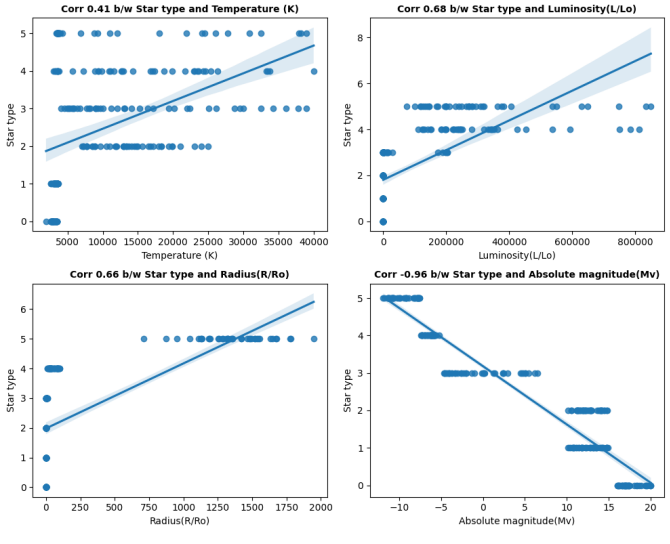


Fig. 9. Bivariate analysis for star types and their features

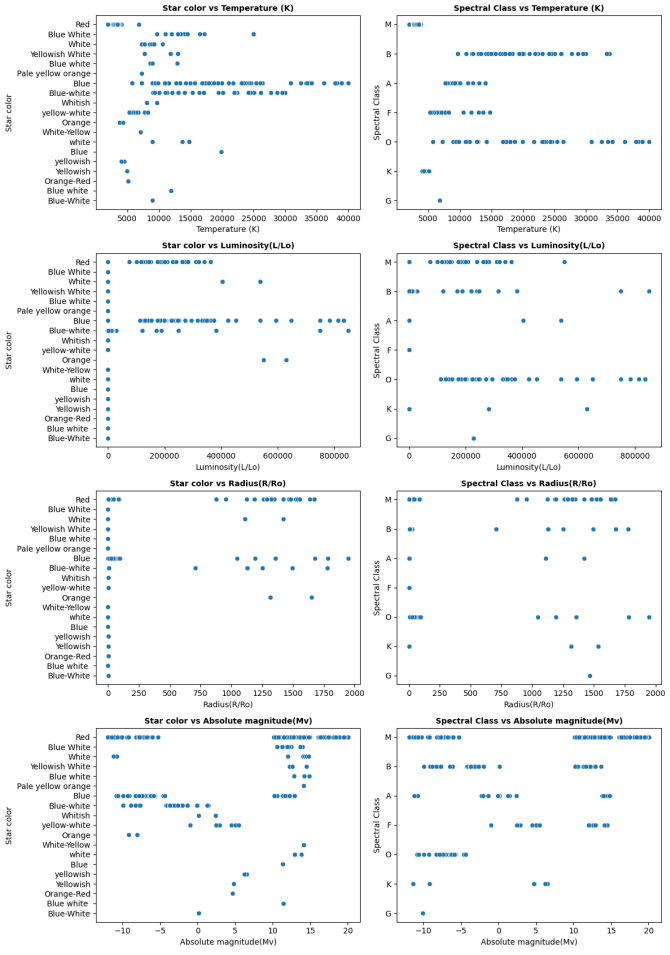


Fig. 10. Temperature, luminosity, radius, absolute magnitude distributions concerning star color and spectral class

using absolute magnitude, which has a logarithmic nature, rather than the previous luminosity values provides better insights into the stars. It should also be noted that, in contrast to previous luminosity values, lower values of absolute magnitude indicate higher luminosity, while higher values indicate lower luminosity.

*7) Box Plots:* Alternatively, box plots can be utilized to see the measurement ranges concerning each star color or spectral class. In Fig. 11 and Fig. 12 the temperature ranges for each star color and spectral class can be seen clearly. The results align with the Fig. 1

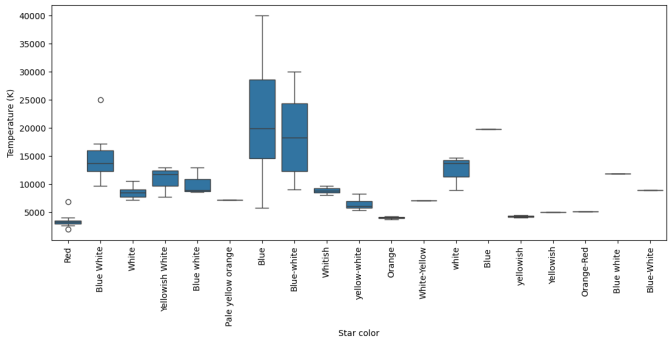


Fig. 11. Box plot for temperature versus star color

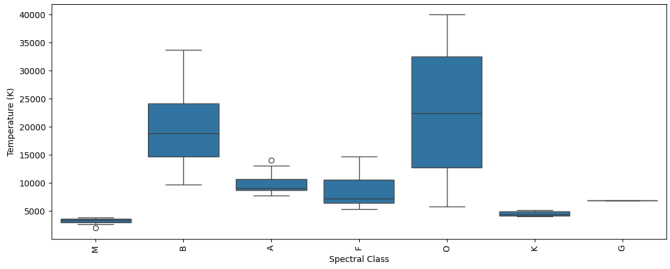


Fig. 12. Box plot for temperature versus spectral class

*8) Multivariate Analysis:* After analyzing the dataset from different perspectives, the general overview of the relationships can be seen in a multivariate analysis given in Fig. 13

Both positive and negative correlations are apparent in Fig. 13. Notably, temperature and luminosity exhibit a strong positive correlation, consistent with the explanation in section III-A6 that absolute magnitude is a representation of luminosity, resulting in a strong negative correlation between temperature and absolute magnitude features. Furthermore, strong correlations are observed between luminosity and radius, absolute magnitude, and star type attributes. It indicates the differentiating impact of luminosity. Focusing on star type, the strongest correlations are observed for luminosity, radius, and absolute magnitudes. This outcome aligns with other analyses discussed throughout the work and is also supported by them.

*9) Correlation Matrix:* A correlation matrix is another technique to demonstrate the correlations between various

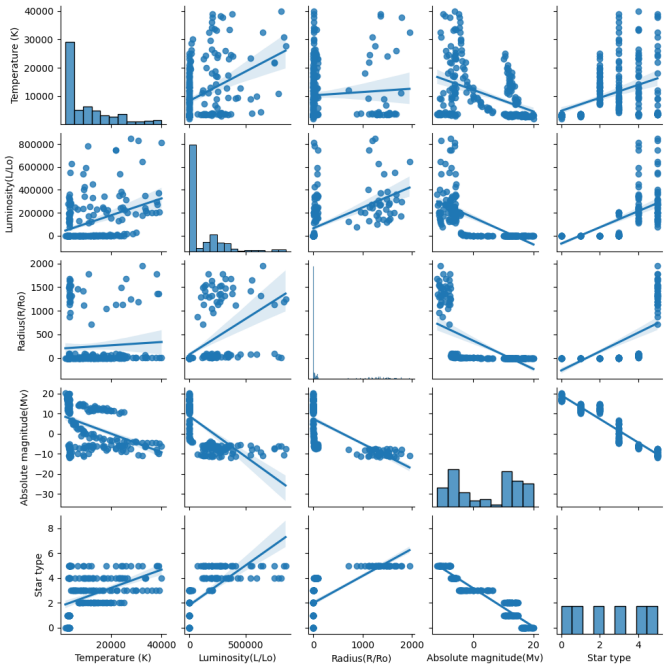


Fig. 13. Multivariate analysis for star features

parameters directly. All the correlation values for each attribute pair in the dataset can be seen in Fig. 14.

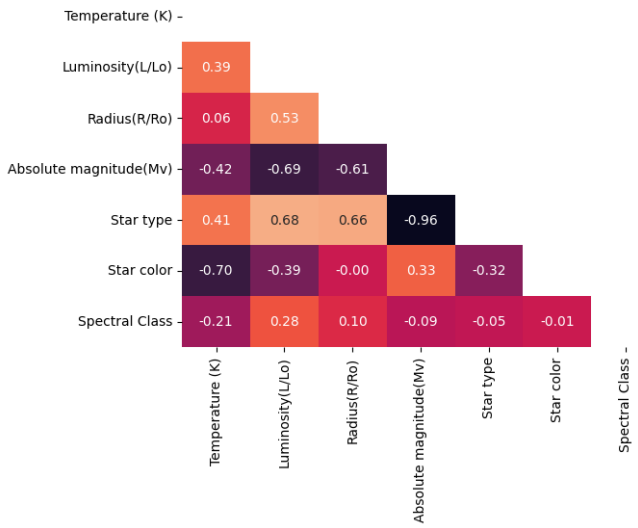


Fig. 14. Correlation matrix for each feature pair

Here, the strongest correlations are observed for star type-absolute magnitude, temperature-star color, and absolute magnitude-luminosity pairs. These high correlations align with the earlier examinations in section III-A6, reinforcing the idea that absolute magnitude provides significant insights about star types. The correlation between star color and temperature is not surprising, given the Fig. 1. The relatively high correlation between absolute magnitude and luminosity is also expected since absolute magnitude is another representation of luminosity. On the contrary, we observe the lowest correlations for star color-radius and spectral class-star color pairs, indicating that assuming star color based on star radius or spectral class based on star color is misleading. Since this work's interest lies in determining star types based on star features, the focus will be on the correlations of these features with star type. These correlations are given in Fig. 15

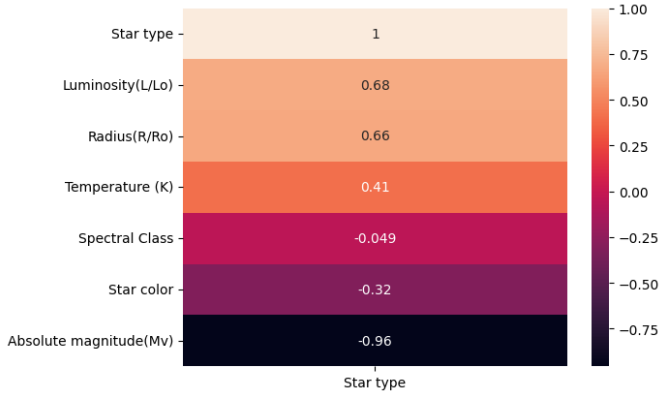


Fig. 15. Heat map for the correlations between the star type and other features

*10) Heat Map:* It can be seen that the absolute magnitude has the strongest negative correlation with the star types. Following it, luminosity has a good correlation with the star type. Since the features are about the star types rather than spectral classes, it is better to examine the correlations with the star type. Because a spectral class may contain wider ranges that make it difficult to find a high correlation. Star color and especially Spectral Class do not have high correlations with the Star type, hence solely depending on this information and this limited dataset in predicting the Star type seems misleading.

## *B. Model Training Results*

As explained in section II-B, Logistic Regression, SVC, K-Nearest Neighbors, Random Forest, and XGB Classifiers are utilized for model training. Initially, the columns that have the lowest correlation in absolute value are dropped alongside the target column. Then all the data is standard scaled to have a zero mean and a standard deviation of 1. This process is also known as Z-value normalization. By this scaling, a more consistent and stable dataset is obtained for model training. After all these preprocessing steps, the results of the models are shown in Fig. 16 and 17.


Fig. 16. Confusion matrices for each algorithm

*1) Confusion Matrices:* Confusion matrices given in 16 indicate all the true and false predictions of the models. In fact, all the models exhibit a good performance in correct predicting but the Logistic Regression has false predictions of Main Sequences for Super Giants. SVC also exhibits the same kind of error. It appears to be a little difficult for models to distinguish Main Sequences from Super Giants. Other models also exhibit some false predictions but in general, they are really good in predicting the star types.

| | Algorithm | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 0 | RandomForestClassifier | 0.966667 | 0.973333 | 0.966667 | 0.966435 |
| 1 | KNeighborsClassifier | 0.950000 | 0.962500 | 0.950000 | 0.950311 |
| 2 | SVC | 0.933333 | 0.953846 | 0.933333 | 0.933333 |
| 3 | XGBClassifier | 0.916667 | 0.922963 | 0.916667 | 0.915931 |
| 4 | LogisticRegression | 0.900000 | 0.918788 | 0.900000 | 0.900791 |

Fig. 17. Score table for the algorithms

*2) Performance Scores:* Accuracy, precision, recall, and F1 scores for each machine learning algorithm are shown in Fig. 17. The meanings of these scores can be briefly explained as follows:

- **Accuracy:** A measure of the overall correctness of predictions.
- **Precision:** A measure of the correct positive predictions. It is equal to the ratio of correct positive predictions to total positive predictions.
- **Recall:** A measure of sensitivity which is equal to the ratio of true positive predictions to the sum of true positive and false negative predictions.
- **F1 Score:** It takes into account both precision and recall to give a better measure. It is defined as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Random Forest Classifier, achieved the highest accuracy score with a well-predictive performance. Also high precision, recall, and F1 scores suggest that the model outperforms other models for various criteria. Also, the confusion matrix shows that the majority of the predictions are correct.

K-Nearest Neighbors Classifier demonstrates a good accuracy right after the Random Forest Classifier and also has consistent precision, recall, and F1 scores. The choice of neighbors and distance might have affected the results. Support Vector Classifier also exhibits similar results with decent and consistent scores.

XGBoost Classifier and Logistic Regression show lower scores than the other models but still quite high results. They demonstrate balanced precision, recall, and F1 scores. Logistic Regression is a simple algorithm and it could show good results.

Apparently all the algorithms show quite high performance on the dataset. Considering the small size of the dataset, these high performances may imply overfitting. Hence, these algorithms' performances should be reassessed with a much larger dataset to get more reliable outcomes.

## IV. Conclusion

In this study, a comprehensive exploration of the Star Types dataset [4] is studied, employing some popular data visualization and machine learning techniques to better understand the relationships between star types and their features. The paper investigated a diverse range of analyses, from multivariate analyses to scatter plots, allowing to give insights into the characteristics of stars.

The dataset is examined in many dimensions regarding the correlations and distributions of key attributes such as temperature, luminosity, radius, and absolute magnitude with themselves and star types. Through many visualizations, patterns that align with the established astronomical knowledge, such as the Hertzsprung-Russell Diagram are shown. Following the data visualizations, machine learning models, including Random Forest, k-Nearest Neighbors, Support Vector Classifier, XGBoost, and Logistic Regression, were utilized to predict star types based on the chosen attributes. The outcomes demonstrated significant accuracy, precision, recall, and F1 scores for all models, highlighting the use of machine learning in stellar classification. Particularly, the Random Forest Classifier emerged as the best model.

While the findings are significant, it is crucial to acknowledge the important limitations of the dataset's size. The

potential overfitting situation should be carefully considered as it reduces the reliability of the findings. A much larger dataset will benefit to validate further and enhance the models.

In conclusion, this study examines the basic concept of using data visualization and machine learning techniques for star classification on a small scale. The data-driven approaches in understanding, modeling, and making predictions in astronomy -in which massive amounts of data are generally observed- seem to accelerate the research about the cosmos further.

## REFERENCES


[1] Swinburne University of Technology, "Harvard Spectral Classification", The SAO Encyclopedia of Astronomy. Accessed: Jan. 3, 2024. [Online]. Available: https://astronomy.swin.edu.au/cosmos/H/Harvard+Spectral+Classification

[2] D. C. Agrawal, 'Apparent and absolute magnitudes of stars: a simple formula', World Scientific News, no. 96, pp. 120–133, 2018.

[3] P. C. Budassi, "HR Diagram", Wikimedia Commons. Accessed: Jan. 3, 2024. [Online]. Available: https://commons.wikimedia.org/wiki/File:HRDiagram.png

[4] D. Baidya, 2019, "Star dataset to predict star types", Kaggle. Accessed: Dec. 10, 2023. [Online]. Available: https://www.kaggle.com/datasets/deepu1109/star-dataset/data

[5] L. Breiman, 'Random forests', Machine learning, vol. 45, pp. 5–32, 2001.

[6] V. Y. Kulkarni and P. K. Sinha, 'Random forest classifiers: a survey and future research directions', Int. J. Adv. Comput, vol. 36, no. 1, pp. 1144–1153, 2013.

[7] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, 'KNN model-based approach in classification', in On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings, 2003, pp. 986–996.

[8] P. Cunningham and S. J. Delany, 'k-Nearest neighbour classifiers-A Tutorial', ACM computing surveys (CSUR), vol. 54, no. 6, pp. 1–25, 2021.

[9] C. Cortes and V. Vapnik, 'Support-vector networks', Machine learning, vol. 20, pp. 273–297, 1995.

[10] T. Chen and C. Guestrin, 'Xgboost: A scalable tree boosting system', in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[11] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, Applied logistic regression, vol. 398. John Wiley & Sons, 2013.

[12] Hughes, D. W., "The Introduction of Absolute Magnitude (1902 - 1922)", ¡i¿Journal of Astronomical History and Heritage¡/i¿, vol. 9, no. 2, pp. 173–179, 2006.

[13] T. Olson, 'The colors of the stars', in Color and Imaging Conference, 1998, vol. 1998, pp. 233–240.