Design of a Classification-Based Machine Learning Model for Star Classification and Hertzsprung-Russell Diagram Development

Parikshit Joshi
DAIICT
202411008@daiict.ac.in

Unique Patel
DAIICT
202411013@daiict.ac.in

Atul Makwana
DAIICT
202411051@daiict.ac.in

***Abstract*—This project aims to design a classification-based machine learning model to classify stars based on their physical properties such as color index, luminosity, temperature, and spectral class. Additionally, the project involves the development of a Hertzsprung-Russell (H-R) diagram to visualize stellar classifications. The project utilizes a dataset of stellar properties and applies data pre-processing techniques. The model is based on Random Forest classification. Results demonstrate a high-accuracy classification model and a visually insightful H-R diagram. Future work includes extending the model for broader datasets and advanced astrophysical interpretations.**

***Index Terms*—Machine Learning, Star Classification, Hertzsprung-Russell Diagram, Data pre-processing, Random Forest.**

I. Introduction

A. Background

Stars are classified based on properties like color index, luminosity, and temperature. The Hertzsprung-Russell diagram is a pivotal tool in astrophysics for understanding stellar evolution. Stars, the luminous celestial bodies, play a crucial role in the dynamics of the universe. Their physical properties such as color index, luminosity, temperature, and spectral class provide valuable insights into their evolutionary stages and classifications. Astronomers have long used the Hertzsprung-Russell (H-R) diagram as a graphical tool to study stars, correlating these properties to identify patterns in stellar evolution. Recent advances in Machine Learning have opened new avenues to automate and enhance the classification of stars based on their physical characteristics, providing an efficient alternative to traditional manual methods.

The H-R diagram serves as the foundation for understanding the life cycle of stars, plotting their absolute magnitudes against surface temperatures to reveal main-sequence stars, giants, supergiants, and white dwarfs. Incorporating ML into this domain can refine star classification by utilizing data-driven models that analyze vast datasets with minimal human intervention, fostering new discoveries in astrophysics.

B. Problem Statement

Classifying stars manually based on their physical properties can be time-consuming and error-prone, especially when dealing with large datasets. Traditional methods rely heavily on human expertise and are limited in scalability and consistency. With the growing availability of astronomical data, there is a pressing need for automated tools that can accurately classify stars and provide insights into their evolutionary stages. The challenge lies in developing a robust classification model that can handle the complexities of stellar data while also visualizing their relationships through the Hertzsprung-Russell diagram.

C. Objectives

- To design and implement a Machine Learning model that classifies stars based on their color index, luminosity, absolute magnitude and spectral class.
- To pre-process and analyze stellar datasets, ensuring data quality and consistency.
- To evaluate the performance of Random Forest classification model.
- To construct the Hertzsprung-Russell diagram for the given dataset, visualizing stellar classifications and relationships.
- To interpret the results to provide meaningful insights into the evolutionary patterns of stars.

D. Scope and Significance

The project leverages Machine Learning techniques to enhance our understanding of stellar classification and evolution. By automating the process, it allows astronomers to handle large-scale datasets efficiently, minimizing human error and enabling faster discoveries. The H-R diagram, augmented with ML-based classifications, provides a modern approach to visualizing and analyzing stellar properties, making it easier to identify trends and anomalies.

The project's significance lies in its interdisciplinary nature, combining astronomy and data science. Its outcomes can contribute to academic research, educational tools, and advancements in astronomical data analysis. Furthermore, the methodology developed can be adapted for other celestial classification problems, promoting broader applications in space exploration and astrophysics.

II. Methodology

A. Dataset

The dataset used for this project includes many key stellar properties. We have taken our dataset from The Astronomy

Nexus. The dataset consists of data of more than 100 thousand stars. It provides insights of about 37 different features of stars ranging from their coordinates, distance, radius, luminosity etc. We have utilized the following features from the dataset:

- Magnitude: The apparent brightness of a star as seen from Earth.
- Absolute Magnitude: The intrinsic brightness of a star, when seen from it at a distance of 10 parsecs.
- Luminosity: Total energy emitted by the star per unit time.
- Spectral Class: Categorical feature representing the classification of stars based on their spectra.
- Color Index: Numerical representation of a star's color, indirectly indicating its temperature. Lower values typically correspond to hotter stars.

B. Pre-processing

1) Handling Missing Values: Since the dataset is huge, it required to pre-process some entries in-order to minimize errors. The dataset consisted of some entries where the values were NaN (Not a Number). Particularly in the Spectral column, the values should all be in Upper case. Some of them were not, and hence they were removed. Rows with missing (NaN) or invalid values in any column were removed.
2) Feature Normalization: Continuous numerical features (mag, absmag, ci) were standardized to ensure consistency in scale.

C. Model Development

1) Train-Test Split: The dataset was split into training and testing sets in an 80:20 ratio using stratified sampling to ensure all spectral classes were adequately represented in both sets.
2) Models Used: Random Forrest Classifier.
3) Evaluation Metrics: Accuracy, Precision, Recall, F1 score, support, macro average and weighted average.

D. H-R Diagram Development

The Hertzsprung-Russell (H-R) diagram was constructed as a scatter plot, showcasing the relationship between luminosity and absolute magnitude of the star. The temperature of a star is a derivative of Luminosity, and is indirectly represented by the color index of star. The key features included are:

- X-axis: Spectral Type and Color Index
- Y-axis: Luminosity and Absolute Magnitude

III. Results and Discussion

A. Classification Model Results

The Random Forest model achieved the highest accuracy of 100%. This high accuracy indicates that the model was able to effectively classify stars based on their properties such as magnitude, absolute magnitude, luminosity, spectral type and color index.

Classification Report:

	precision	recall	f1-score	support
A	1.00	1.00	1.00	3743
B	1.00	1.00	1.00	2051
C	0.97	1.00	0.98	29
D	1.00	0.97	0.98	32
F	1.00	1.00	1.00	5146
G	1.00	1.00	1.00	4491
K	1.00	1.00	1.00	6372
M	1.00	1.00	1.00	1026
N	1.00	1.00	1.00	11
O	1.00	1.00	1.00	49
R	0.93	0.93	0.93	14
S	0.75	1.00	0.86	3
W	1.00	0.93	0.96	14
accuracy			1.00	22981
macro avg	0.97	0.99	0.98	22981
weighted avg	1.00	1.00	1.00	22981

Fig. 1. Classification report

Below is the confusion matrix for our model, which provides detailed insights about the model's accuracy.

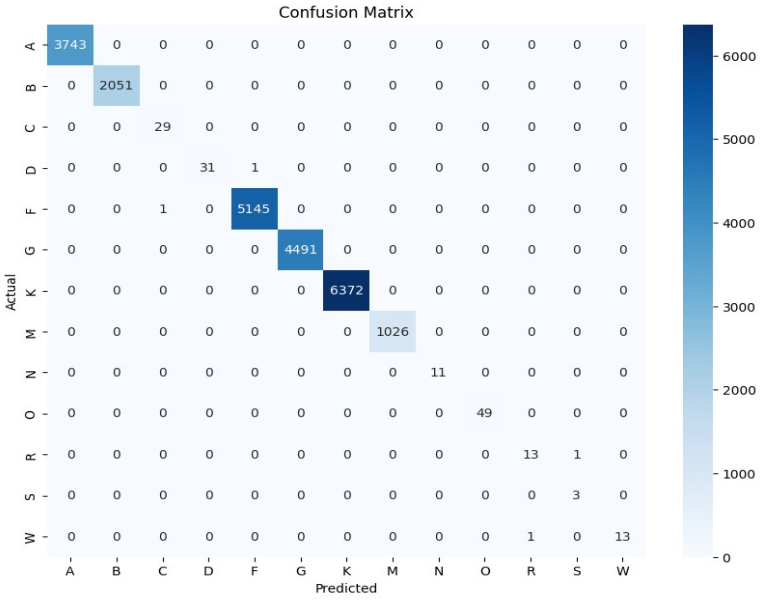

Fig. 2. Confusion matrix diagram

B. H-R Diagram

Figure 3 shows the Hertzsprung-Russell (H-R) diagram, which plots the relationship between absolute magnitude, luminosity, spectral type and color index.

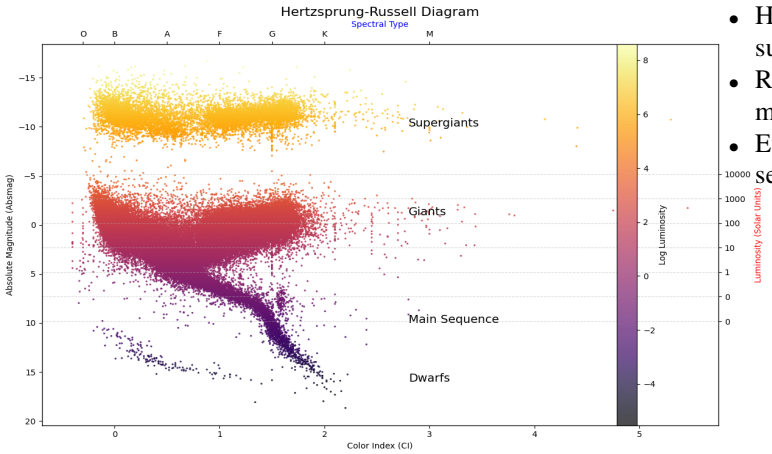

Fig. 3. Hertzsprung-Russell Diagram

IV. Conclusion and Future Work

A. Conclusion

The project successfully classified stars based on their properties using machine learning models. The Random Forest model achieved the highest accuracy of 100%, demonstrating the effectiveness of ensemble methods in star classification. Additionally, the Hertzsprung-Russell (H-R) diagram was developed to visualize the relationship between stellar luminosity and temperature, providing valuable insights into stellar evolution.

B. Future Work

Future improvements to this project could include:

- **Enhance Model Performance with Deep Learning:** Deep learning models could be explored to further improve classification accuracy.
- **Incorporate Additional Stellar Parameters:** Additional features such as stellar age, metallicity, and radial velocity could be integrated.

References

- The Astronomy Nexus: https://www.astronexus.com/
- Scikit-learn Documentation, Available online at https://scikit-learn.org/
- Matplotlib Documentation, Available online at https://matplotlib.org/
- Wikipedia-The Hertzsprung-Russel diagram https://en.wikipedia.org/wiki/Hertzsprung-Russell_diagram
- Database: The Astronomy Nexus (GitHub Page): https://github.com/astronexus?tab=repositories/

Appendix

- Color index: A numerical value that indicates its temperature and color (b-v).
- Luminosity: Total amount of energy star emits per second across all wavelengths.
- Spectral class: A classification system that describes the star's surface temperature and ionization state based on the strength of its spectral length.
- Hertzsprung-Russell diagram (H-R Diagram): Plots the surface temperature of stars against their luminosity
- Random Forest: A machine learning algorithm that uses multiple decision trees to make predictions.
- Evaluation Metrics: Quantitative measures that helps assess how well a machine learning model is performing.