Data-Driven Spectral Classification of Stars with Machine Learning

Hanzhi Chen

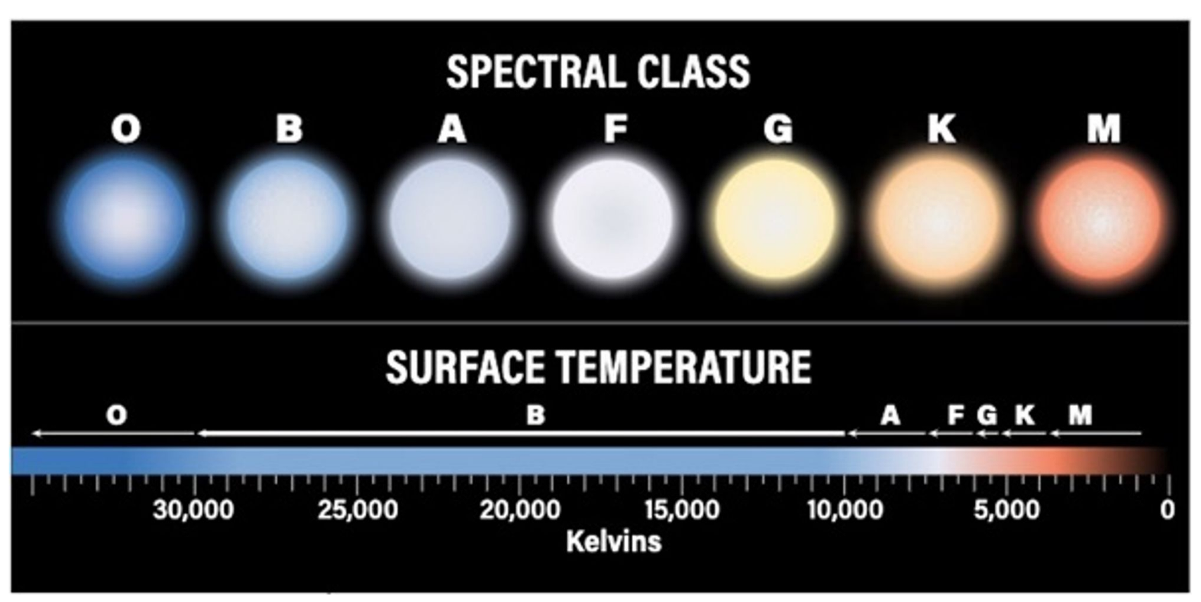
Mentor: Ryan Jackim

Abstract

We present a study using machine learning and feature engineering techniques, including column transformation and recursive feature elimination (RFE), for the multi-class classification of stars' spectral types using their photometry information. Our proposed supervised learning approach (Random Forest and Support Vector Machine) was trained on labeled stellar datasets. Two models were developed and compared to evaluate their performance. The experimental results demonstrated that the proposed method with feature engineering improved the classification accuracy and computational efficiency of the models. This research demonstrates the potential of using machine learning algorithms for improving the accuracy and efficiency of star classification tasks, which could be particularly beneficial for quicker classification of a large number of objects compared to traditional classification methods.

Background

- Spectral classification is key in astronomy for understanding star properties like temperature and chemical composition.
- Machine learning offers a promising alternative to traditional subjective and time-consuming methods
- Large-scale astronomical datasets from Sloan Digital Sky Survey (SDSS) and Gaia eDR3 have enabled the training of machine learning models.



[1] The Harvard spectral classification system classify stars based on their spectral lines. It puts stars into seven main types: O, B, A, F, G, K, and M, where O-type stars are the hottest and M-type stars are the coolest.

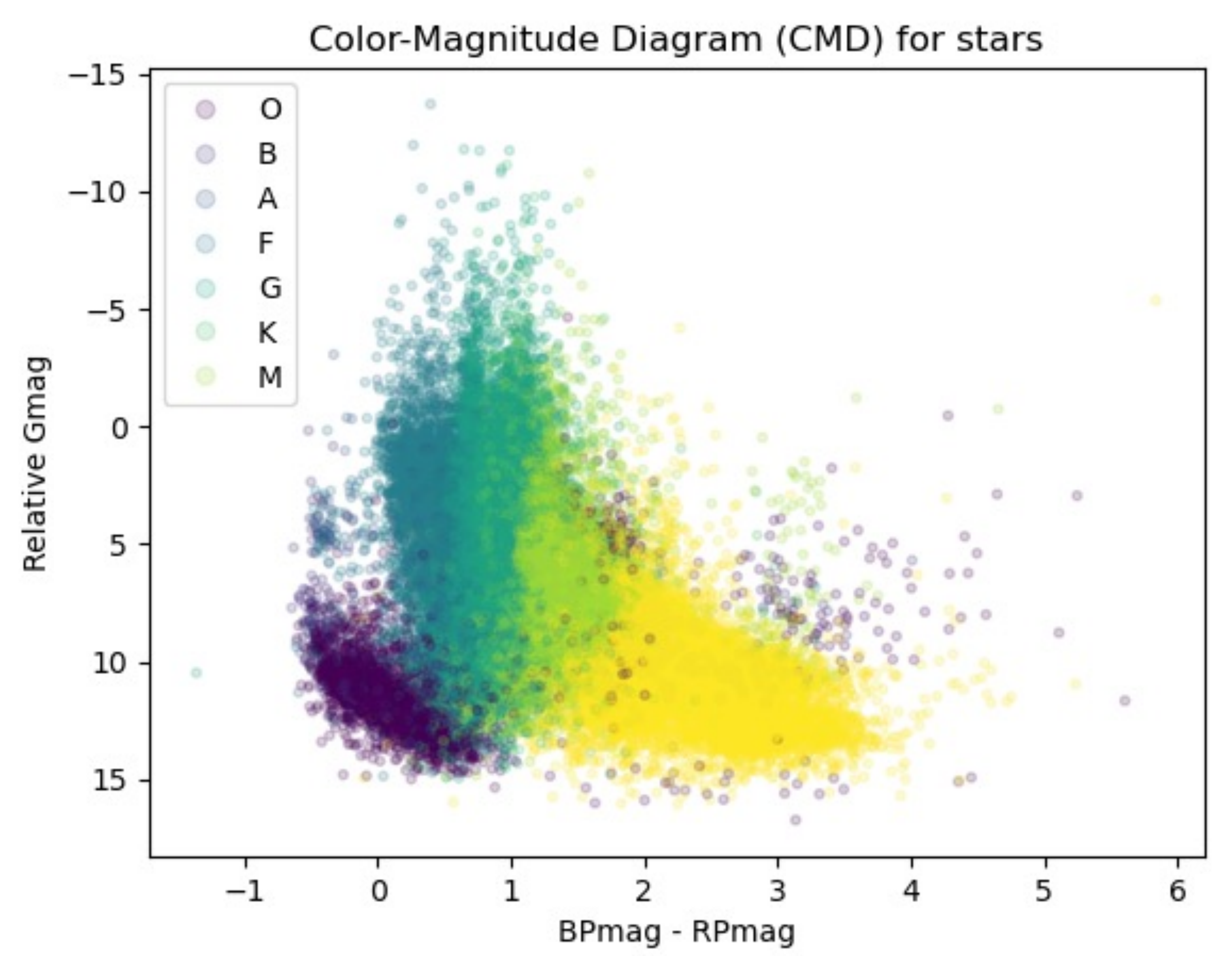
Goal

- Use feature engineering and machine learning models to classify stars into corresponding spectral class.
- Train and cross-validate the model on datasets from SDSS and Gaia eDR3 mainly using photometry data.
- Evaluate the performance of models with confusion matrices.

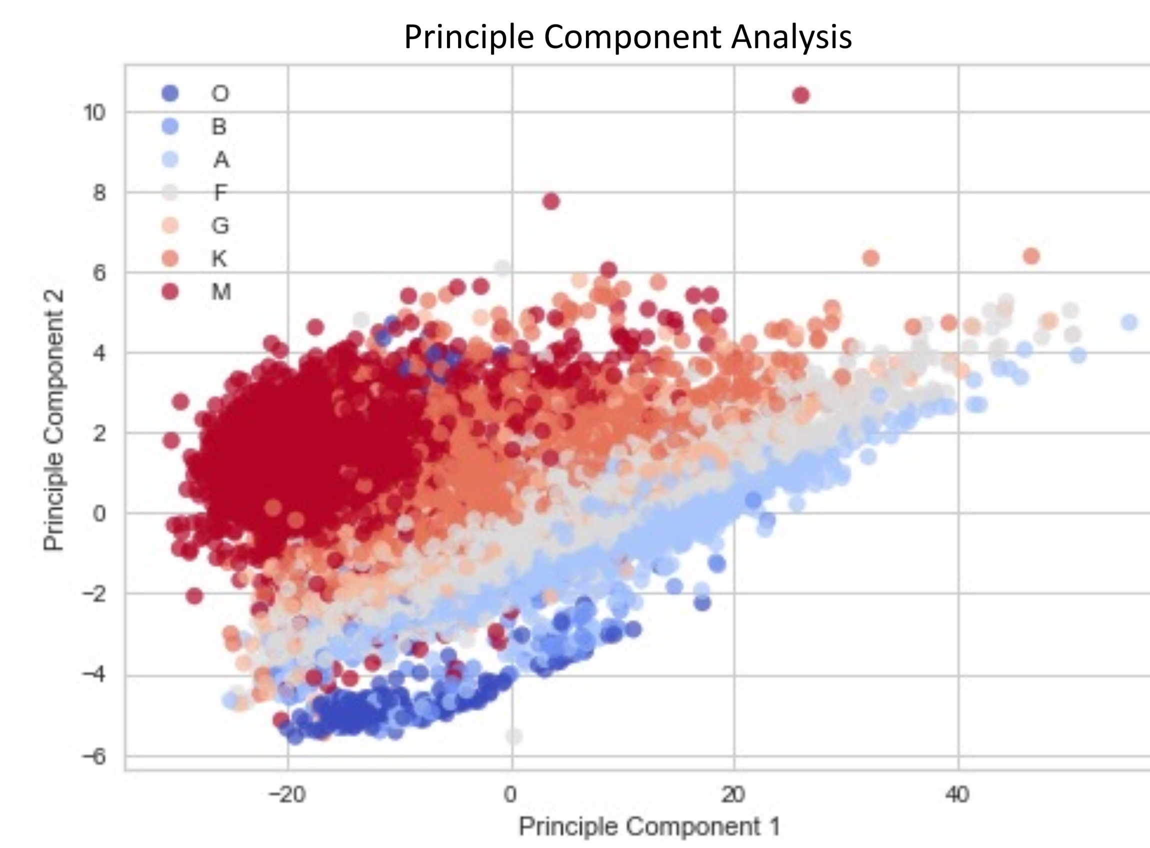
Methods

Data Analysis and Feature Engineering

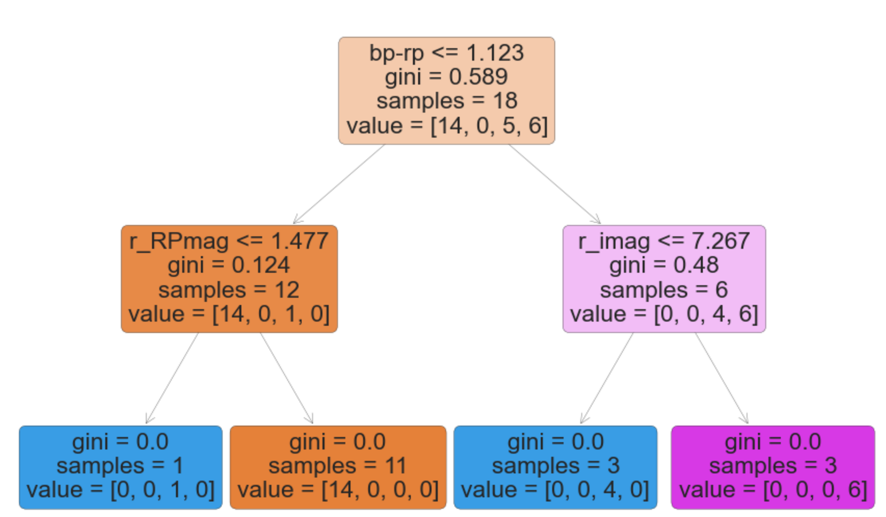
- Visualized data in 2D with Principle Component Analysis (PCA) to show the existence of a relationship between the star's magnitude and its spectral type.
- Selected features that contribute the most to the prediction by conducting Recursive Feature Elimination to reduce the computational complexity and storage requirements.
- Obtained a dataset of 100,000 astronomical objects classified as stars from the SDSS and cross-matched it with the Gaia eDR3 dataset to extract additional features.
- Converted the absolute magnitude to relative magnitude to for distance correction using parallactic shifts from Gaia eDR3.
- Calculated the difference in magnitude between the blue and red filters, which is commonly used as a measure of the color of the star for Gaia eDR3, for improved classification.



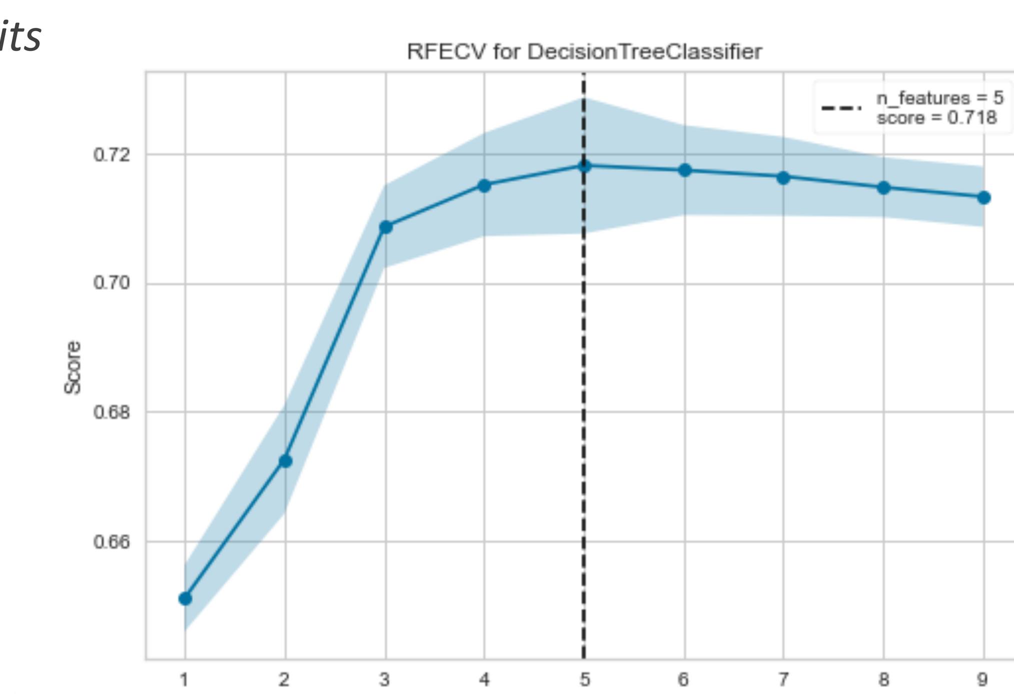
By plotting relative Gmag against BPmag-RPmag, we explore the relationship between a star's color and its brightness (or magnitude).



PCA reduces the dimensionality of the data by identifying the most important features that contribute to the variance.



An example of one decision tree in a random forest model trained with 25 samples. Bigger dataset yields more complex trees which are difficult to visualize.



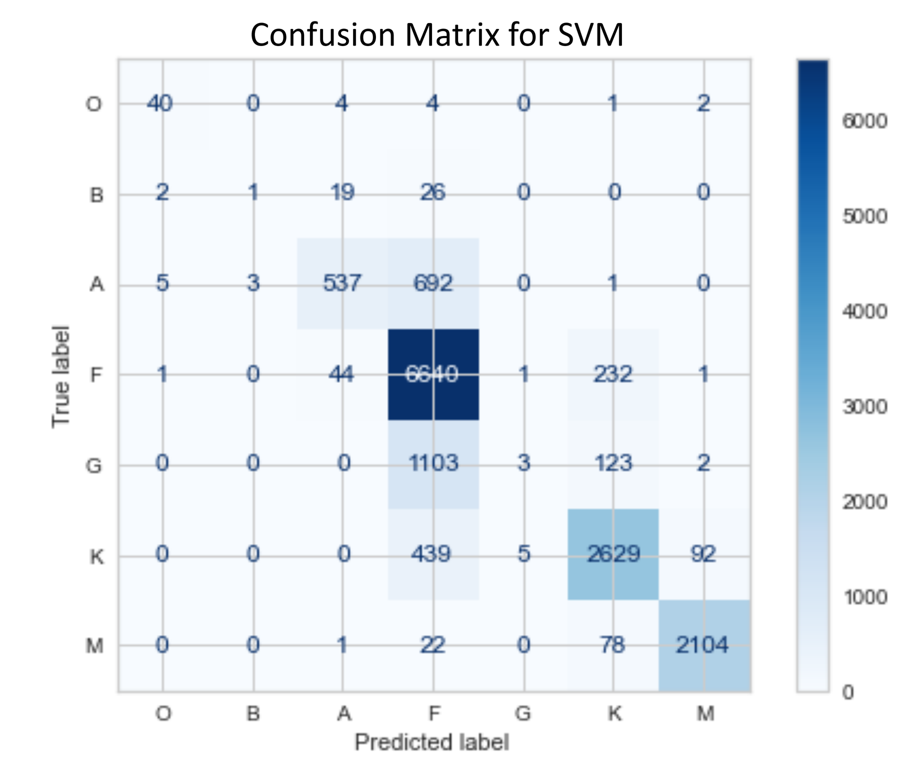
Recursive feature selection showed that the decision tree classifier performs best with only 5 features, as the validation score increases with the number of features until it reaches a maximum at 5 and then gradually decreases.

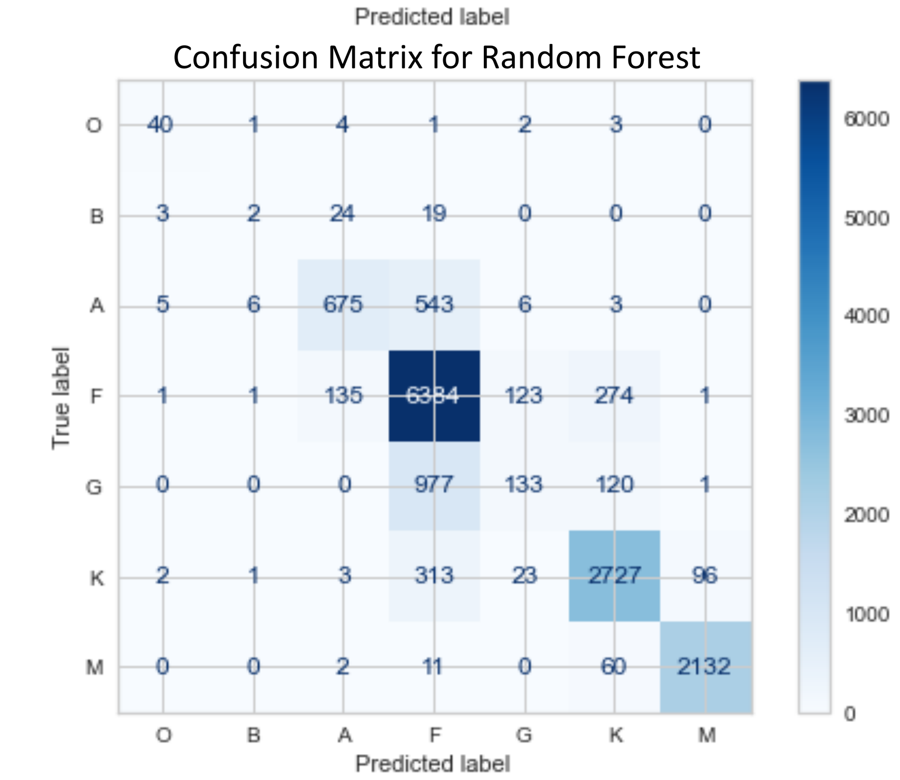
Models

- Support Vector Machine (SVM) with RBF kernel — uses random search for optimal hyperparameters.
- Random Forest — uses an ensemble of 100 decision trees to make predictions

Results

- Both Random Forest and SVM models achieved accuracy of about 80%.
- SVM with RBF kernel had slightly better performance (82.12%) compared to random forest (81.57%), but random forest was faster to train.
- Star type distribution did not fully align with the theoretical expectation, possibly due to the small training set.





Future Work

- Scale up the training dataset to improve generalization ability
- Expand features to other wavelengths in the electromagnetic spectrum.
- Apply deep learning techniques to model directly on raw spectral data

Reference:

1. Roen, K. (2021, April 20). *Color-coding stars*. Astronomy.com. Retrieved February 23, 2023, from https://astronomy.com/magazine/glenn-chaple/2021/04/color-coding-stars