Data Analysis

The study leverages key data science libraries to process a 240-star dataset with 6 mixed-type features for predictive modeling of spectral classes. Despite the imbalance favoring class M, the aim is to accurately classify stars. Analysis included histograms and scatter plots, revealing diverse distributions and a notable radius-luminosity correlation.

Data Preprocessing and Handling Class Imbalance

The dataset initially contained a single sample of the spectral class 'G'. Machine learning models require sufficient samples for each class to learn patterns effectively. Having only one instance in the 'G' class would skew the model's learning process and evaluation metrics, making it unreliable. Therefore, the single sample of spectral class 'G' was removed from the dataset to avoid these issues. Categorical features such as Star color and Spectral Class were encoded into numerical values using LabelEncoder. Next, we standardized the features in the dataset. The features varied in scale, such as temperature, luminosity, and radius. K-means clustering and neural network algorithms like MLP and CNN are sensitive to the scale of input data. Standardizing features ensures that each feature contributes equally to the model's calculations. StandardScaler was used to transform the features so that they have a mean of 0 and a standard deviation of 1. The dataset exhibited a significant class imbalance, which could lead to biased models that favor majority classes. To address this, we used class weights and the SMOTE technique.

In supervised learning, imbalanced datasets can cause models to perform poorly on minority classes. Using class weights in the loss function helps the model to give more attention to the minority classes during training. The weights were set inversely proportional to the class frequencies, penalizing misclassifications of minority classes more heavily. The Synthetic Minority Over-Sampling Technique (SMOTE) was employed to generate synthetic samples for minority classes, creating a more balanced training dataset. This approach is particularly useful for training neural networks, which require sufficient data to learn effectively. SMOTE was applied to the training data, ensuring that the MLP and CNN models received a more uniform distribution of class samples during training, leading to better generalization.

Model Implementation

Four Models were implemented to classify the stars: Random Forest, K-means clustering, MLP, and CNN.

Random Forest model, which is an ensemble of decision trees, was used as a baseline supervised learning method. The model was trained using the standardized features, with class weights applied to handle class imbalance. The K-means algorithm was initialized with 6 clusters, matching the number of spectral classes. Features were standardized before clustering, and the resulting clusters were mapped to the actual classes using majority voting for evaluation. The MLP and CNN models were adapted for structured data classification in which the input data was reshaped consisting of two hidden and two convolutional layers respectively and the output layer had 6 neurons for the spectral classes. SMOTE was applied to balance the training data to handle class imbalance. The model was trained with some parameters and evaluated on the test set.