**Executive Summary**

This project aimed to analyze a cosmology dataset containing various characteristics of stars, such as temperature, luminosity, radius, and absolute magnitude, to identify anomalies that could represent unusual or rare celestial objects. Employing unsupervised learning models, including the Isolation Forest and Local Outlier Factor (LOF), we developed methodologies capable of detecting anomalies with distinct characteristics. The analysis revealed stars that deviate significantly from typical patterns, potentially indicating new types of celestial phenomena. Recommendations include further astronomical validation of identified anomalies and the exploration of additional features to enhance anomaly detection capabilities.

**Introduction**

The field of astrophysics continually seeks innovative methods to discover and classify celestial objects. This project focuses on detecting anomalies within a dataset of stellar properties, aiming to uncover unique or rare stars that deviate from common patterns. By applying unsupervised learning techniques, specifically Isolation Forest and LOF models, we aimed to identify such anomalies, contributing to the broader understanding of the cosmos and supporting the discovery of new astronomical phenomena.

**Data Overview**

The dataset comprises various stellar characteristics, including temperature, luminosity, radius, and absolute magnitude, alongside categorical variables such as star color and spectral class. Preprocessing steps involved scaling numerical features and encoding categorical variables to prepare the data for anomaly detection analysis.

**Methodology**

*Data Exploration:* Initial analysis focused on understanding the distribution of stellar characteristics and identifying preliminary patterns within the data.

*Preprocessing:* Included scaling of numerical features to standardize the data, facilitating effective anomaly detection by the models.

*Anomaly Detection Models*: Utilized Isolation Forest and LOF models for their proficiency in identifying outliers within complex datasets. The models were chosen for their ability to handle the dataset's mixture of numerical and categorical features effectively.

*Evaluation:* Employed visualization techniques and silhouette scores to assess the effectiveness of the anomaly detection models, alongside domain-specific evaluation criteria.

*Results:*

The anomaly detection models successfully identified stars with unusual properties, with the Isolation Forest model showing a higher silhouette score compared to the LOF model, suggesting a more distinct separation of anomalies. Analysis of the identified anomalies revealed stars with significantly higher temperature, luminosity, and radius, potentially indicating rare types of celestial objects. Cross-validation with astronomical catalogs and further spectral analysis are recommended to validate these findings.

**Recommendations**

*Astronomical Validation:* Further investigation and cross-referencing of identified anomalies with astronomical catalogs to confirm their uniqueness and potential significance.

*Feature Exploration:* Exploration of additional stellar properties and the incorporation of spectral data to enhance the models' ability to detect a broader range of anomalies.

*Model Refinement:* Continued refinement of anomaly detection models, including parameter tuning and the exploration of alternative unsupervised learning techniques, to improve detection accuracy and interpretability.

**References**

<https://www.kaggle.com/code/ybifoundation/stars-classification>