Project Proposal

Title: Predicting Stellar Characteristics from Hipparcos Star Catalog Data Using Data Mining Techniques and Machine Learning

Team Members:

1. Vishnu Pradhaban Jawahar (C836879478)
2. Vidyavathi Kumar (C837235421)
3. Tanisha Mehta (C836877056)

1. Brief Description of the Problem/Opportunity

The Hipparcos Star Astronomy Catalog s a collection of data from the Hipparcos satellite, which was a space-based astronomy mission launched by the European Space Agency (ESA) in 1989.
This catalog provides a rich dataset of over 118,000 stars, including essential properties such as position, brightness, and motion. The challenge lies in effectively utilizing this dataset to classify stars based on their brightness levels and other characteristics. Such classification can enhance our understanding of stellar populations, aiding astronomers and researchers in their investigations into stellar characteristics and behaviors.

2. Specific Business Objective(s) of Your Analysis

The primary objective of this analysis is to enhance the understanding of stellar populations and support astronomical research by developing predictive models that classify stars. Specifically, the focus is on identifying variable stars (whether a star is a variable star or not – binary classification task) based on their astrometric and photometric data.
Note: A variable star is a star whose brightness changes over time due to intrinsic or extrinsic factors. These variations in brightness can occur over periods ranging from minutes to years, depending on the type of variable star.

3. Predictive Modeling Task(s)

The modeling task will focus on classifying stars into binary categories, specifically predicting whether a star is classified as a variable star (yes/no) based on various attributes such as brightness, distance, and color indices.

4. Potential Dataset(s) and Sources

- **Dataset Name:** Hipparcos Star Catalog

- **Source:** Kaggle
- **Link:** Hipparcos Star Catalog
- **Approximate Number of Observations:** 118,000 stars

Please note that we will be performing data cleaning while preparing the dataset for modeling. We expect the number of observations to be reduced by up to 30%.

5. Training and Validation Split

- **Training Observations:** 82,600 (70%)
- **Validation Observations:** 35,400 (30%)

Note: These numbers are tentative, but we will maintain a 70% split for training and 30% for validation throughout the analysis.

6. Potential Dependent Variable

- **Dependent Variable:** Stellar classification as a variable star (binary: yes/no)

7. Potential Independent Variables

- **Independent Variables:**
  - Vmag (Visual magnitude)
  - Plx (Parallax)
  - RAdeg (Right Ascension in degrees)
  - DEdeg (Declination in degrees)
  - B-V (Color index)

Note: Additional variables may be considered and included as part of the final model, which will be determined through further analysis and modeling techniques.

8. Data Mining Techniques

We will employ the following data mining techniques:

- **Decision Trees:** To classify stars based on their characteristics, generating a visual representation of the classification process.
- **Logistic Regression:** To predict the likelihood of a star being a variable star based on independent variables.

- **Neural Networks:** To explore advanced classification techniques for identifying variable stars.

9. Data Mining Software

The analysis will be conducted using:

- **KNIME**
- **Python Libraries:** scikit-learn, TensorFlow (specific choice of libraries may vary based on analysis requirements)

Conclusion

This project aims to leverage the comprehensive Hipparcos Star Catalog to develop predictive models that classify stellar characteristics, thereby enhancing the field of astronomical research. By employing various data mining techniques, we hope to contribute valuable insights into the nature of stars and their behaviors.