**Classification of Star Clusters using GAIA DR3**

**DATA 301: Individual Project 2023**  
Sakila Wanigasinghe (300526406)  
Other Group Members: Alif Nurrokhman, Theodore Loach

Table of Contents

Background

The Gaia mission is a space observatory mission launched by the European Space Agency (ESA) with the primary goal of mapping the Milky Way galaxy in unprecedented detail such as observing the positions, motions, and other properties of over a billion stars.

Leveraging the GAIA DR3 data, the aim of this research is to accurately classify stars belonging to the same star clusters within a specified region of the sky. Star clusters, groupings of stars born from the same interstellar cloud, provide a unique glimpse into the processes of star formation, evolution, and the dynamics of the galaxy. By extracting and analysing relevant parameters such as stellar positions, proper motions, and photometric properties from GAIA DR3, this research aims to develop sophisticated machine learning and data analysis techniques that enable the precise classification of stars that share a common origin within these clusters.

GAIA DR3 is a vast dataset containing data on approximately 1.8 billion different sources. To conduct this project, subsets of the larger GAIA DR3 data were used, using 8,000 sources. Only 2000 records can be pulled from a ADQL query at a time, due to data limitations set by the ESA, which limited how much data could be extracted in this particular use case.

The GAIA DR3 data does not contain star cluster labels, so a secondary dataset containing was used for cross-matching to create a labelled dataset suitable for classification model training.

**Star clusters of interest**

To determine whether a classification model can be used to determine which stars are a member of the same clusters, a few clusters in particular have been selected to explore.

The four clusters considered in the exploratory data analysis are:

* NGC 2437 (Messier 46)
* NGC 6819 (Foxhead Cluster)
* NGC 7789 (Caroline’s Rose Open Cluster)
* Trumpler 5

NGC 2437, also known as Messier 46, is an open star cluster located in the Puppis constellation. This cluster is situated in the southern celestial hemisphere and is recognised for its rich population of stars. Messier 46 contains over 500 stars, with a prominent red giant at its centre. It is estimated to be around 250 million years old and is positioned against a backdrop of fainter stars and interstellar dust.

NGC 6819, often referred to as the Foxhead Cluster, is an open star cluster situated in the constellation Cygnus. This cluster is relatively young, with an estimated age of around 2 billion years, making it a fascinating subject for stellar evolution studies. NGC 6819 is named for its distinctive star arrangement, which, when observed visually, resembles the shape of a fox's head. It contains a diverse population of stars, from hot and blue to cooler and redder stars.

NGC 7789, also known as Caroline's Rose Open Cluster, is in the constellation Cassiopeia. This open star cluster is notable for its resemblance to a delicate rose when viewed through a telescope, earning it the nickname. With an estimated age of approximately 1.7 billion years, NGC 7789 hosts a multitude of stars in varying stages of their lifecycle, from young, hot stars to aging giants.

Trumpler 5 is a star cluster found within the constellation Carina. It is one of the well-studied clusters in the region and is categorised as an open cluster. The cluster is recognised for its notable collection of bright stars, some of which are prominent O-type and B-type stars, known for their high temperatures and luminosities.

These star clusters will be used to train and evaluate the performance of a classification model which aims to correctly identify stars as being a part of a certain cluster (or not).

Data Description

In the GAIA DR3 data, there are over 150 different features available. As a result, eleven features were selected, including a unique source identifier (twelfth feature). These eleven features contain raw parametric data which are used to create many of the additional derived features in the GAIA DR3 data. These features are easier to interpret than the more complex derived features. Additionally, it is important to note that the feature *‘teff\_val’*, which refers to the effective temperature of a star, is not available in the DR3 release (it was available in the DR2 release). Instead *‘teff\_gspphot’* which was not available in the DR2 release, is available in the DR3 release.

Four separate labelled datasets were created using these same features, each for a different star cluster.

*Table 1* lists the features and their descriptions that were extracted from GAIA DR3 using ADQL queries in Python, and the additional class (target) variable from the secondary dataset, found in each dataset.

|  |  |
| --- | --- |
| Feature Name | Feature Description |
| source\_id | Unique identifier for each source |
| ra | Right ascension astronomical coordinate in degrees |
| dec | Declination astronomical coordinate in degrees |
| parallax | Angle used to approximate distance from Earth in milliarcseconds |
| pmra | Proper motion in the right ascension direction in milliarcseconds per year |
| pmdec | Proper motion in the declination direction in milliarcseconds per year |
| phot\_g\_mean\_mag | Mean magnitude in the G band (apparent magnitude) |
| phot\_bp\_mean\_mag | Mean magnitude in the BP band |
| phot\_rp\_mean\_mag | Mean magnitude in the RP band |
| bp\_rp | BP-RP index (colour, difference in magnitudes) |
| radial\_velocity | Radial velocity of star compared to the  background |
| teff\_gspphot | Effective temperature which is estimated  from the BP-RP |
| Cluster | Target variable, states whether a star is in the desired cluster or not |

*Table 1: Features in each labelled dataset used for classification model training.*

Each feature in *Table 1* is numerical, except *‘Cluster’*, which is nominal categorical. Out of the numerical features, *‘ra’* and *‘dec’* together provide the coordinates for a star. All the numerical variables are numerical continuous. *‘Cluster’* contains two levels, one for if the star is in the specified cluster, and the other for if it is not.   
  
**Class balance in each dataset**

|  |  |  |
| --- | --- | --- |
| Dataset (Star Cluster) | Instances in Cluster | Instances not in Cluster |
| NGC 2437 | 1166 | 834 |
| NGC 6819 | 1024 | 872 |
| NGC 7789 | 1165 | 837 |
| Trumpler 5 | 547 | 1453 |

*Table 2: Class balance in each dataset.*

*‘Table 2’* illustrates the class balance (or imbalance) in each of the datasets. In all datasets except the one pertaining to the Trumpler 5 cluster, there are more positive instances (instances in the cluster) than negative instances (instances not in the cluster). In these cases, the class imbalance is not overly huge.

However, in the case of the Trumpler 5 dataset, there are only 547 instances that are a part of the cluster out of 2000. This indicates a considerable class imbalance which may need to be remedied using synthetic sampling and/or class weightings.

**Missing data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature Name | NGC 2437 | NGC 6819 | NGC 7789 | Trumpler 5 |
| phot\_bp\_mean\_mag | 16 | 30 | 9 | 20 |
| phot\_rp\_mean\_mag | 16 | 28 | 9 | 20 |
| bp\_rp | 16 | 30 | 9 | 20 |
| radial\_velocity | 1791 | 1755 | 1606 | 1871 |
| teff\_gspphot | 176 | 102 | 119 | 196 |

*Table 3: Features with missing data (and how many missing instances), for each cluster dataset.*

*Table 3* highlights that *‘radial\_velocity’* across all four datasets is afflicted with lots of missing data. Considering that most of the data is missing for that column, using that feature as a predictor may not be suitable. *‘teff\_gspphot’* has between 100 and 200 missing values across each of the datasets. This is a smaller proportion, which can be imputed accurately. The same applies to the other three features as well.

Ethics and Privacy

The ethical and privacy considerations pertinent to the utilisation of the GAIA DR3 data are intrinsically linked to the specialised nature of this research, focused on the classification of star clusters. The dataset exclusively engages with celestial objects in space, and therefore inherently avoids the typical ethical concerns associated with personal privacy, copyright, or fair use.

It is imperative that this research rigorously adheres to the data usage policies and guidelines set forth by the ESA and the GAIA mission. Moreover, the ethical importance of proper acknowledgment and citation of the ESA and the GAIA mission in any research publications, not only as an ethical obligation but also as a demonstration of scientific integrity is required. In addition to these ethical principles, it is important to consider the ethical value of open science and collaboration.

Like ethical concerns, the usual privacy concerns relating to most types of data do not necessarily apply to the GAIA DR3 data. An important thing to consider is that although the data is not personal, it is important to ensure that other sources indirectly connected to individuals or organisations are protected from disclosure.

Furthermore, there is a cultural element to be mindful of, specifically if astronomic work was to broach constellations and bodies of special significance to a particular culture. The means in which results would be discussed and communicated should consider the cultural importance it may have to certain peoples. Ideally, the best approach would be to communicate with concerned members and parties where appropriate and possible.

Project data and results will be kept secure through a combination of measures. Implementing a robust data backup strategy and routinely duplicating project data to secure offsite locations will ensure that data is useful. This practice not only safeguards against data loss but also ensures data availability in case of unforeseen events. Data will be stored on secure servers and cloud platforms, benefiting from encryption and stringent access controls. To further bolster security, regularly monitoring and maintaining access logs, tracking who accesses project data, when, and what actions are taken, enables to prompt identification and response to any suspicious activity. These precautions are integral to maintaining the confidentiality and integrity of the project data and research findings.

Exploratory Data Analysis (EDA)

To understand how the eleven numerical features interact with each other, three types of plots were produced (for each cluster dataset). Triangle correlation plots were created to identify features that are correlated. Declination vs. right ascension scatter plots were produced to understand how the stars look in space, and see where outliers, determined by parallax (stars not in a cluster) may lie. Lastly, colour-magnitude diagrams were



