**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. Initially, ADQL queries were used to extract data. However, only 2000 records can be pulled from a ADQL query at a time, due to data limitations set by the ESA. As a result, using SQL queries allowed for more data to be extracted.

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).

Star clusters, when viewed from Earth as constellations, are often grouped together with similar coordinates (right ascension and declination, defined in the coming sections). However, even though in the sky they look to be a part of the same group, in space these stars vary in how far away they are from Earth. Stars that are further away from Earth than other stars are generally not in the same cluster.

Currently, there is no direct measurement which measures how far a star is from Earth, hence why this project aims to construct a classification model to identify stars within in the same cluster.

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. The datasets have differing numbers of rows, as this depends on the number of relevant GAIA data available for each cluster.

*Table 1* lists the features and their descriptions that were extracted from GAIA DR3 using SQL 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 | Total Instances |
| NGC 2437 | 1231 | 946 | 2177 |
| NGC 6819 | 1024 | 872 | 1896 |
| NGC 7789 | 2060 | 1515 | 3575 |
| Trumpler 5 | 829 | 2473 | 3302 |

*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 829 instances that are a cluster member out of 3302 total stars. 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 | 17 | 30 | 16 | 24 |
| phot\_rp\_mean\_mag | 17 | 28 | 16 | 24 |
| bp\_rp | 17 | 30 | 16 | 24 |
| radial\_velocity | 1948 | 1755 | 2902 | 3075 |
| teff\_gspphot | 188 | 102 | 228 | 304 |

*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 300 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 purpose of classifying stars within 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.

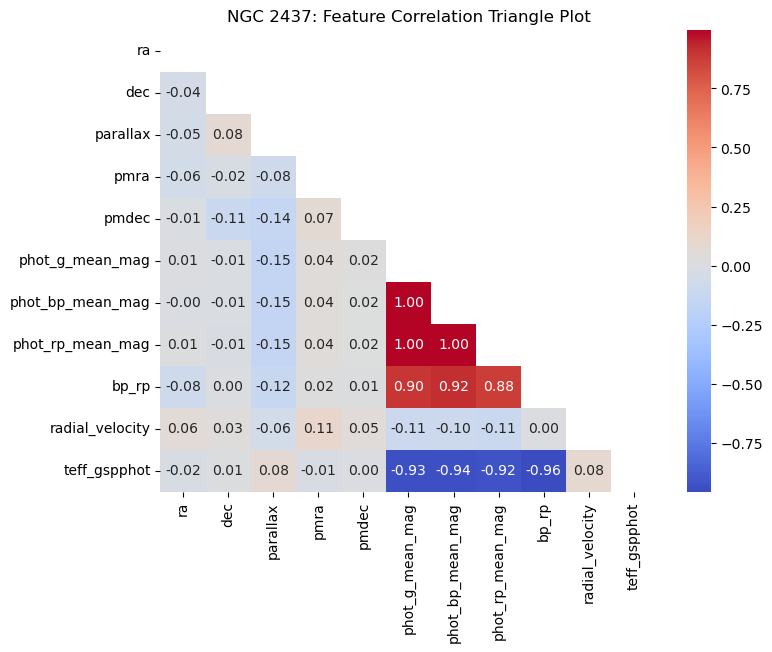
Project data and results will be kept secure through a combination of measures. Implementing a robust data backup strategy and routinely duplicating project data to secure offsite locations will ensure that data is safe. This practice not only safeguards against data loss but also ensures data availability 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 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 extract some initial insights from the data, and understand any unusual patterns which may need to be considered, three key plots were produced for each cluster, along with a more general data quality assessment:

* Feature correlation triangle plots
  + Highlights any strong correlations between features.
* Right ascension vs. declination, by parallax scatter plots
  + A two-dimensional representation of stars in space, useful for observing stars that vastly differ in parallax to the rest.
* Colour-magnitude diagrams
  + Understand the temperature of individual stars in relation to its intrinsic brightness (absolute magnitude).

Three plots were produced for each cluster as a part of the analysis, however, not all these plots have been included in this report, as some of them conveyed the same, if not very similar information to some of the other plots.  
  
**Feature Correlations**

Among all four clusters, the highest correlations were between the three features that provide information about the observed brightness of a star. *‘phot\_g\_mean\_mag’*, *‘phot\_bp\_mean\_mag’*, and *‘phot\_rp\_mean\_mag’* each had extremely strong, positive correlations with one another, ranging between 0.96 and 1.00. It is logical that these features are highly correlated, considering they all pertain to the observed brightness of a star.

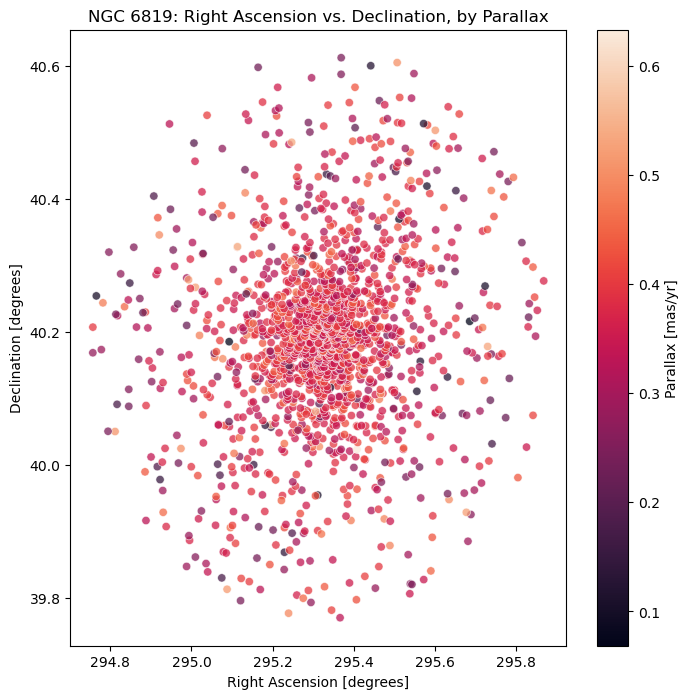
For the NGC 2437 cluster, there were also very strong positive correlations between each of the three features listed above, and *‘bp\_rp’*. *‘bp­\_rp’* is the difference in magnitudes between *‘phot\_bp\_mean\_mag’*, and *‘phot\_rp\_mean\_mag’*, hence as a derived feature it makes sense it is highly correlated. Interestingly, the other three star clusters did not exhibit strong correlations between *‘bp­­\_rp’* and the three other features.

*Figure 1: Feature correlations in the NGC 2437 cluster.*

The other correlations of note across all four datasets were the medium to strong negative correlations between *‘teff\_gspphot’* and the four features. The NGC 2437 and NGC 7789 clusters had the strongest correlation, and the other clusters had weaker correlations. Again, the correlation is sensical considering that *‘teff\_gspphot’* is estimated from *‘bp\_rp’*. Furthermore, this indicates that the effective temperature of stars increases as brightness does, in particular photometric bands.

**Right ascension vs. declination, by parallax**

Right ascension and declination are used to specify the position of celestial objects on the celestial sphere, similar to longitude and latitude on Earth. Parallax, on the other hand, is a measurement technique used to determine the distances to nearby stars by taking advantage of the apparent shifts in their positions as Earth orbits the Sun.

Stars with extreme parallax values compared to the other stars with similar right ascension and declination values are often not a part of a cluster.

Out of the four clusters, NGC 6819 had a considerable number of stars with significantly lower parallax values compared to the other stars in the same region, as illustrated in *Figure 2*.

*Figure 2* highlights a handful of stars with parallax values at approximately 0.1 mas/yr. These points are identified as they are darker in colour and differ from the majority of the stars which have a parallax somewhere between 0.3 and 0.5 mas/yr. Compared to other clusters, NGC 6819 had the most range in parallax values amongst stars with similar right ascension and declination values.

*Figure 2: Right ascension vs. declination for NGC 6819*

In attempting to produce a similar plot for NGC 7789, it became clear there was erroneous values in the data. 264 instances in the data contained extremely different right ascension values. These 264 instances had right ascension values approximately between 0 and 1 degrees, whilst the rest of the instances had right ascension values approximately between 358 and 360 degrees. The correct right ascension of stars in the NGC 7789 cluster was ascertained using secondary sources. The data was then consequently filtered to exclude the 264 erroneous instances.

**Colour-magnitude diagrams**

Right ascension and declination are used to specify the position of celestial objects on the celestial sphere, si