**Classification of Star Cluster Members 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 in a common region of space, as being a cluster member or not.

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, SQL queries were used to allow for more data to be extracted.

The GAIA DR3 data does not contain star cluster labels, so a secondary dataset containing star cluster labels 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, two clusters have been selected to explore:

* NGC 7789 (Caroline’s Rose Open Cluster)
* Trumpler 5

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 in the same region of space. 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 a similar region of space that are members of a particular 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 region. Each dataset contains information about individual stars in the same region of space, as observed from Earth. Not all these stars are a part of the 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 Region) | Instances in Cluster | Instances not in Cluster | Total Instances |
| NGC 7789 | 2060 | 1515 | 3575 |
| Trumpler 5 | 829 | 2473 | 3302 |

*Table 2: Class balance in each dataset.*

*‘Table 2’* illustrates the class imbalance in each of the datasets. The Trumpler 5 region data is severely imbalanced, indicating the need for synthetic sampling, class weighting adjustments, or other imbalance remediation.

**Missing data**

|  |  |  |
| --- | --- | --- |
| Feature Name | NGC 7789 | Trumpler 5 |
| phot\_bp\_mean\_mag | 16 | 24 |
| phot\_rp\_mean\_mag | 16 | 24 |
| bp\_rp | 16 | 24 |
| radial\_velocity | 2902 | 3075 |
| teff\_gspphot | 228 | 304 |

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

*Table 3* highlights that *‘radial\_velocity’* across both 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 200 and 300 missing values across the datasets. This is a smaller proportion, which can be imputed accurately. The same applies for 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 cluster members within a certain region of space. 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.

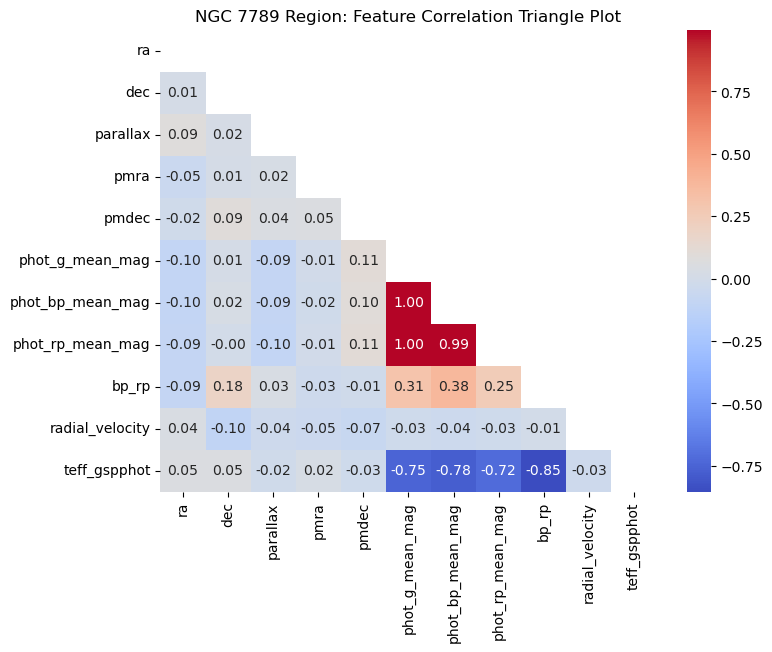
Project data and results will be kept secure through a combination of measures. Implementing a robust data backup strategy and routinely duplicating project data to secure offsite locations will ensure that data is safe. This practice not only safeguards against data loss but also ensures data availability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

* 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).

In total, six plots were produced (three for each cluster region), however not all of these plots have been included as they did not all reveal interesting or notable insights.  
  
**Feature Correlations**

Among both clusters regions, 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.

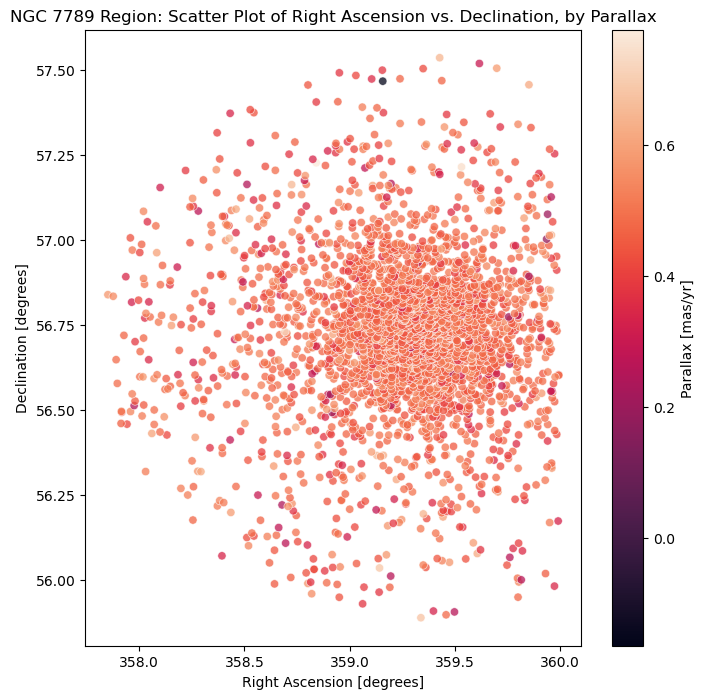
The other correlations of note were present in the NGC 7789 dataset. There were medium to strong negative correlations between *‘teff\_gspphot’* and the four features mentioned previously. Again, the correlation is sensical considering that *‘teff\_gspphot’* is estimated from *‘bp\_rp’*.

*Figure 1: NGC 7789 cluster region feature correlations.*

Furthermore, this indicates that the effective temperature of stars increases as brightness does, in particular photometric bands. These correlations were not present in the Trumpler 5 cluster region data.

**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.

*Figure 2* highlights a handful of stars with parallax values at less than 0.3 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 above 0.4 mas/yr.

In attempting to produce this plot for the NGC 7789 region, it became clear there was erroneous values in the data.

*Figure 2: NGC 7789 cluster region visualised.*

264 instances in the data contained extremely different right ascension values. These 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 region was ascertained using secondary sources. The data was then consequently filtered to exclude the 264 erroneous instances.

A similar plot was produced for the Trumpler 5 region data, however, there were no points were extreme parallax values compared to the rest. In general, the parallax of stars in this region were a lot lower, ranging between -0.2 to 0.2.

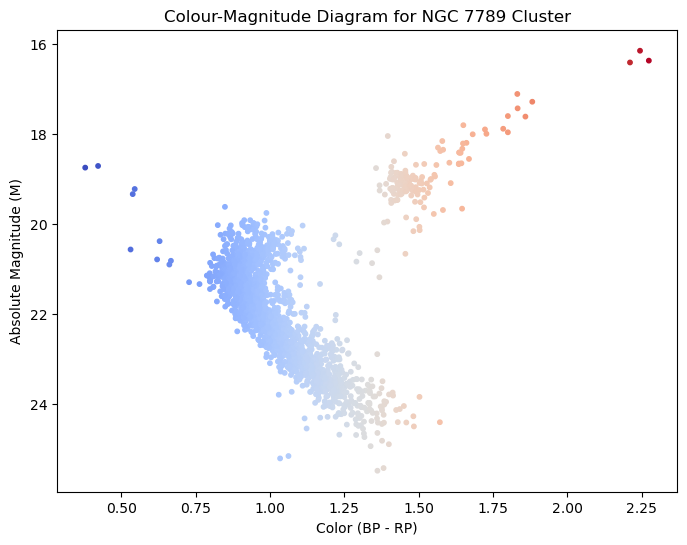
**Colour-magnitude diagrams (CMD)**

The absolute magnitude of a star (derived using *‘phot\_g\_mean\_mag’* and *‘parallax’*) is a measure of the luminosity of a star.

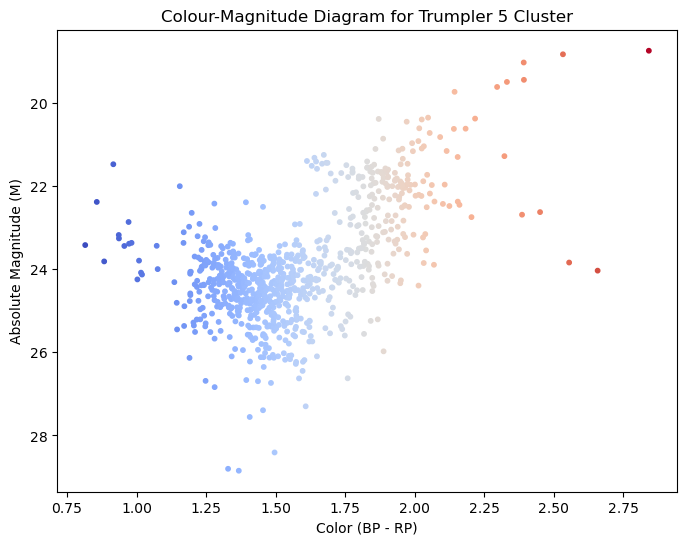
Using absolute magnitude and *‘bp\_rp’,* colour-magnitude diagrams can be produced to understand which stars are in the ‘prime’ of their lives and the age of a cluster. Clusters are generally groups of stars that were formed at approximately the same time. The CMD diagrams produced only contain stars which are cluster members, as the aim of these graphs is to understand the age of the star clusters, and their unique features.

The ‘main sequence’ in a CMD is the prominent diagonal feature that extends from the top left (bright, blue stars) to the bottom right (dim, red stars). The main sequence consists of stars that are in the prime of their lives and are actively fusing hydrogen into helium in their cores.

The ‘turn-off’ point is the point where the main sequence starts to turn off to the right (toward higher "bp-rp" values). Younger clusters have turn-off points located to the left (bluer) of the main sequence, while older clusters have turn-off points farther to the right (redder).



*Figure 3: NGC 7789 cluster colour-magnitude diagram.*



*Figure 4: Trumpler 5 cluster colour-magnitude diagram.*

*Figure 3* illustrates a clear turn-off point for stars in the NGC 7789 cluster region, along with some blue ‘stragglers’ just to the left and above the main sequence. There is notable a gap in space between stars on the main sequence and stars turning off the main sequence. Initially, this was assumed to be indicative of an incomplete dataset, however, further research into the NGC 7789 cluster, including comparisons to other CMDs have revealed that this is not a data issue, but a unique feature of the cluster **(include referencing, to do later).**

The turn-off point is farther to the right where stars are redder, indicating that the cluster is older.

*Figure 4* does not depict a straightforward main sequence like *Figure 3*, which is indicative of an incomplete dataset. However, the CMD for the Trumpler 5 cluster still provides useful insights. Also, unlike the NGC 7789 cluster, the turnoff point is not overly disjoint from the main sequence. This is since Trumpler 5 is a considerably old cluster, meaning many stars have already transitioned to the later stages of their evolution. Consequently, the brightest and bluest stars are less numerous, contributing to the less well-defined main sequence.

<https://www.researchgate.net/figure/CMD-of-NGC-7789-Filled-squares-denotes-stars-with-membership-probability-P-75-open_fig3_2219669>

<https://deepeye.hu/en/drawings/ngc7789.html>

Detailed Analysis Results

hi