Stargazing

Searching for open clusters in GAIA data

Taine Murphy

Team Members:
Van Vo
Max Tan

2022-11-03

Acknowledgement:

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular, the institutions participating in the Gaia Multilateral Agreement.

1. Executive Summary

This report explores the data provided by the GAIA satellite, and applies the clustering method DBSCAN to find stars in open clusters. Given an area of sky containing a known cluster, the algorithm is able to detect this cluster accurately. Additionally, it was able to detect two known clusters within the same area of the sky. The model achieved F1 scores ranging from 55% to just under 90%, with an average score of 71.5%.

From the Exploratory Data Analysis (EDA), we understood the completeness of the datasets of the open clusters M67 and Pleiades. Looking at the data cuts we were able to highlight three important features of cluster membership, being the proper motions and parallax of the stars. After this, the class labels were applied to the dataset, which was acquired from an external dataset in order to have a performance metric.

The DBSCAN model was tuned for the single-cluster and two-cluster subsets of data, and the optimum values were used in the final models. DBSCAN is a density-based clustering algorithm, which is able to filter out large amounts of ‘noisy’ data by finding closely-related data points.

The work done hopefully sets a foundation to work on discovering new open clusters in space and learning more about what an open cluster is.

2. Background

How can we use stellar properties of measurements from the GAIA satellite to determine open-cluster membership, by using the clustering method DBSCAN?

This project has made use of the GAIA data set produced by the European Space Agency, from mission GAIA (https://www.cosmos.esa.int/gaia). Mission GAIA is a satellite launched to record billions of objects in the sky, with the purpose to gain a deeper understanding of the history and evolution of the Galaxy.

This satellite records information on the location, proper motions, parallax, colour, magnitude, etc., of billion celestial objects, from which a three-dimensional mapping of the Galaxy is formed. The GAIA instruments and payload setup allow the European Space Agency (ESA) to detect even the faintest objects and observe them a few times a year at different positions in the sky. With such information, GAIA has the power to unveil the Galaxy's history and also predict its future.

Fig.1: *Image of the Pleiades open cluster, where the Matariki Constellation is located.*

Open clusters are small groups of stars that are loosely bound and tend to be irregular in shape. There are numerous open clusters in the sky, and many more are thought to continue to be discovered. The inexact science of classifying an open cluster has been well documented, and machine learning is a tool that can be used to help find out more about the stars in our skies.

3. Data Description

There are a total of 153 columns within the GAIA dataset “gaiadr3.gaia_source”.

The range of data types includes:
• char
• short
• float
• boolean
• double

This covers a range of information about the dataset including, but not limited to:
• Unique identifiers
• Position
• Movement (direction, speed)
• Distance
• Photometry (colour, brightness)
• Correlation between features
• Classification probabilities (Quasar vs Galaxy vs Star)
• Measuring metrics (eg. measurement error)

3.1. Data Gathering/Sampling Method

We ran two queries to extract two subsets of the dataset, each representing two open clusters (Pleiades and m67). This was extracted using the GAIA archive which filtered based on location, proper motion and error rates. These were collated using multiple different sources to construct an ADQL query within the archive to extract these subsets (Cánovas, 2022; Heyl et al., 2022).

For the Detailed Analysis, this was extended further to include 3 more clusters (M7, Ruprecht 171 and NGC 6645). For the modelling aspect of the analysis, a performance metric was required in order to assess the modelling capabilities. A dataset published on VizieR by Cantat-Gaudin & Anders (2020) was used, which applies a classification algorithm UPMASK to assign membership probabilities to stars (Krone-Martins & Moitinho 2014; Cantat-Gaudin et al. 2018). An example query can be found in the Appendices (Appendix A), where I extracted only the source IDs of those stars with cluster membership of 70% and above. With these IDs, I combined them with the original dataset GAIA to form the class labels.

So overall, there are 3 single-cluster subsets (M7, M67 and Pleiades), and one double-cluster subset (Ruprecht 171 and NGC 6645). The double-cluster subset was an additional test for the model, where it clusters two known open clusters within the same sky area.

3.2. Feature Selection

The features chosen in this project are ‘source_id’, ‘ra’, ‘dec’, ‘pmra’, ‘pmdec’, ‘parallax’, ‘bp_rp’, ‘phot_g_mean_mag’ and ‘teff_gspphot’.

Source ID is the unique identifier for a source within the GAIA dataset.

RA and Dec stand for Right Ascension and Declination. They are astronomical measures of the location of stars in the sky, similar to the latitude and longitude that are used for the earth’s surface. Right Ascension corresponds to east/west direction (like longitude), while Declination measures north/south directions (like latitude). These features are measured in degrees.

Proper motion is an astronomical measure of the movement of stars in the plane of the sky. Therefore, PMRA and PMDec are the proper motions in the directions of Right Ascension and Declination. This is measured in milliarcseconds per year.

Parallax is the apparent displacement of an object because of a change in the observer’s point of view. Astronomers can use insights derived from the parallax measurements of the closer stars to estimate distances of those more distant. This is measured in milliarcseconds.

BP-RP is the colour of each star. More specifically, how blue it is. The more blue a star is, the higher its temperature. This is calculated by subtracting the mean magnitude of the RP filter from the mean magnitude of the BP filter. This is measured in magnitude.

BP-RP colour: phot_bp_mean_mag - phot_rp_mean_mag

Phot_g_mean_mag is the mean magnitude in the G band, which represents the luminosity/perceived brightness of the source. This is computed from the G-band mean flux applying the magnitude zero-point in the Vega scale and is measured in magnitude.

Teff_gspphot is the effective temperature which is estimated from the BP-RP, apparent magnitude and parallax in a process called GSP-Phot (Andrae, 2022). It is measured in Kelvin (k).

3.3. Data Completeness

The right ascension and declination features exist for all records in the GAIA dataset. The queries from the GAIA archive were constructed in a way to filter out missing values of the proper motion and parallax. There were some missing data in terms of luminosity and temperature in all of the subsets, but these were only used for visual purposes (Colour-Magnitude diagram), so it wasn’t necessary to be handled for the model.

4. Ethics, Privacy and Security

One of the most important ethical considerations we emphasise is properly acknowledging the authorship of the GAIA data in our project. Authorship and accreditation are ultimately important, especially in scientific research, as they proclaim the contribution one has made to broaden our understanding of a specific matter. Thus, we want to ensure ethical sourcing of the GAIA data by giving credit to the people and institutes that have worked hard to make the GAIA data accessible to the public. Therefore, an acknowledgement is given to the European Space Agency, and there are citations of their work.

Additionally, preventing a wrong understanding of the GAIA data is ethical. This raises the question of whether the data and methods used can produce reliable results. According to the ESA, there are a few prevailing issues with the data set, most commonly seen as insufficient data (many missing values). Having this in mind helps us determine a better approach to the data by focusing on relevant variables that have plenty of entries.

There are few privacy and confidentiality concerns, as the data is not about people and has also been published by the data owner (ESA) for public scientific usage of the data.

5. Exploratory Data Analysis

5.1. Summary Tables

Pleiades

```
RangeIndex: 15118 entries, 0 to 15117
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   source_id        15118 non-null  int64
 1   ra               15118 non-null  float64
 2   dec              15118 non-null  float64
 3   pmra             15118 non-null  float64
 4   pmdec            15118 non-null  float64
 5   parallax         15118 non-null  float64
 6   phot_g_mean_mag  15118 non-null  float64
 7   bp_rp            15048 non-null  float64
 8   teff_gspphot     13662 non-null  float64
```

M67

```
RangeIndex: 13152 entries, 0 to 13151
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   source_id        13152 non-null  int64
 1   ra               13152 non-null  float64
 2   dec              13152 non-null  float64
 3   pmra             13152 non-null  float64
 4   pmdec            13152 non-null  float64
 5   parallax         13152 non-null  float64
 6   phot_g_mean_mag  13149 non-null  float64
 7   bp_rp            13079 non-null  float64
 8   teff_gspphot     11494 non-null  float64
```

Pleiades

	source_id	ra	dec	pmra	pmdec	parallax	phot_g_mean_mag	bp_rp	teff_gspphot
count	1.511800e+04	15118.00	15118.00	15118.00	15118.00	15118.00	15118.00	15048.00	13662.00
mean	6.670932e+16	56.67	24.14	7.06	-11.95	2.23	15.47	1.59	4868.84
std	1.834031e+15	1.09	1.00	18.72	20.49	2.46	1.96	0.69	1003.69
min	6.396231e+16	54.42	22.12	-293.34	-1157.43	-17.06	3.62	-0.40	2679.21
25%	6.508193e+16	55.80	23.34	-0.81	-14.21	0.85	14.47	1.08	4194.84
50%	6.659966e+16	56.70	24.17	4.27	-6.74	1.34	15.68	1.34	4945.43
75%	6.824963e+16	57.55	24.95	11.34	-2.69	2.54	16.78	1.98	5574.06
max	7.023251e+16	58.79	26.11	611.11	196.93	74.99	20.47	4.68	13439.77

M67

	source_id	ra	dec	pmra	pmdec	parallax	phot_g_mean_mag	bp_rp	teff_gspphot
count	1.315200e+04	13152.00	13152.00	13152.00	13152.00	13152.00	13149.00	13079.00	11494.00
mean	6.032847e+17	132.81	11.80	-6.75	-7.97	1.81	15.47	1.38	4896.59
std	3.558064e+15	0.98	0.96	16.38	17.39	2.07	1.88	0.66	917.81
min	5.976414e+17	130.81	9.82	-322.84	-387.85	-9.36	4.21	-0.42	2845.99
25%	5.988853e+17	132.06	11.06	-11.07	-10.77	0.85	14.34	0.86	4270.03
50%	6.046911e+17	132.82	11.80	-5.65	-4.23	1.20	15.62	1.13	4974.42
75%	6.051195e+17	133.54	12.54	-0.90	-1.21	2.10	16.79	1.79	5626.49
max	6.086413e+17	134.88	13.81	163.07	111.49	85.80	20.24	4.75	18292.03

Fig.2: *Full data frame information collected directly from the GAIA archive for both the Pleiades and M67 open clusters. The top two tables show the counts and data types of variables, while the two bottom tables show the summary statistics of the variables.*

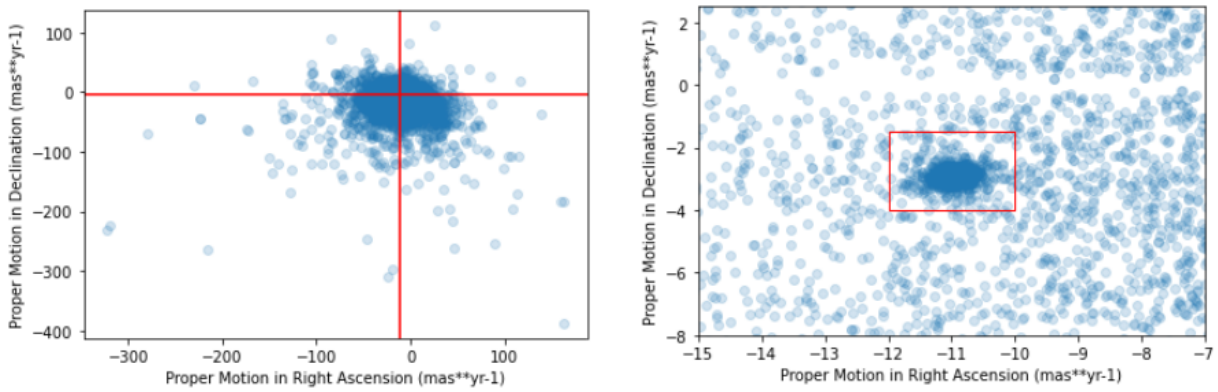
The two datasets have 15,118 and 13,152 records respectively, with both missing data in colour (bp_rp) and temperature (teff_gspphot), and the M67 dataset missing some data on luminosity (phot_g_mean_mag). These values are only needed for plotting purposes (for example the Colour-Magnitude diagram), so no handling of the values is required.

Between the main variables (ra, dec, pmra, pmdec, parallax) there is a large difference in mean and standard deviation, so standardisation and scaling are required in future analysis.

5.2. Data Filtering

To locate the clusters within the subsets, the proper motions and parallax were used to make data cuts and inspect the data (Cánovas, 2022; Heyl et al., 2022). Information from the SIMBAD astronomical database on the proper motions was found so that the cluster can be easily found.

M67



Pleiades

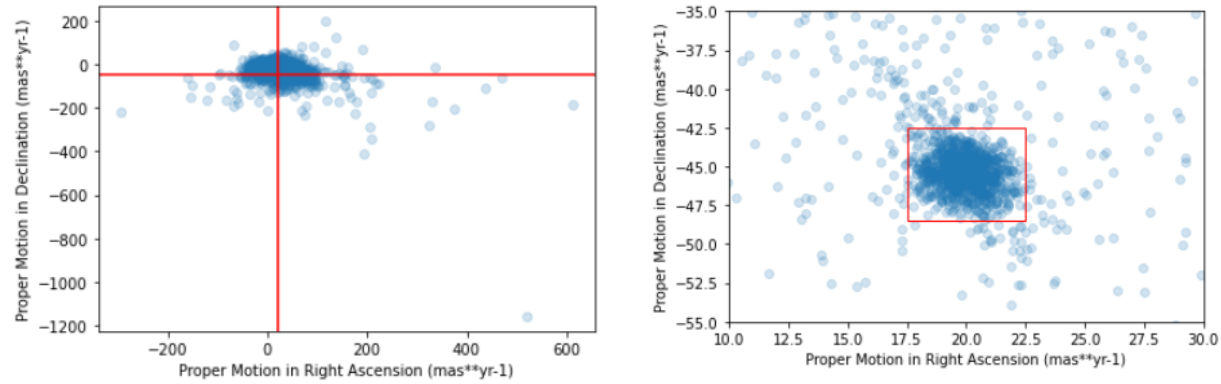
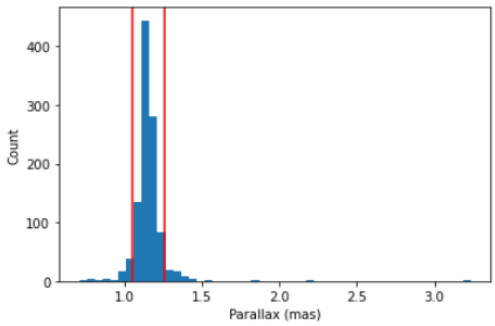


Fig.3: *Proper motion plots. Each left plot displays the overall subset, with the red cross indicating the centre of the cluster. Each right plot displays a zoomed-in view of the centre of the cluster, with a red square indicating the location of the cluster formation.*

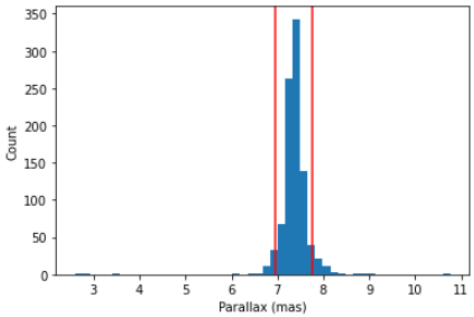
Proper motion is an astronomical measure of the movement of stars in the plane of the sky. Therefore, PMRA and PMDec are the proper motions in the directions of Right Ascension and Declination. This is measured in milliarcseconds per year.

M67




```
count    1062.000
mean        1.158
std         0.106
min         0.709
25%         1.124
50%         1.153
75%         1.184
max         3.234
Name: parallax, dtype: float64
```

Pleiades




```
count    945.000
mean       7.362
std        0.398
min        2.588
25%        7.249
50%        7.370
75%        7.484
max       10.763
Name: parallax, dtype: float64
```

Fig.4: *Parallax distributions after the filter on the proper motion. Red lines show 1 s.d from the mean. The tables on the right show the summary statistics of the parallax from these histograms.*

Parallax is the apparent displacement of an object because of a change in the observer's point of view. This is measured in milliarcseconds.

Both the proper motions and parallax features have similar values to each other, as shown in the plots above (Fig.3, Fig.4). This indicates an opportunity to utilise clustering algorithms to detect open clusters. A density-based algorithm such as DBSCAN allows for finding these small but densely populated sections of large datasets.

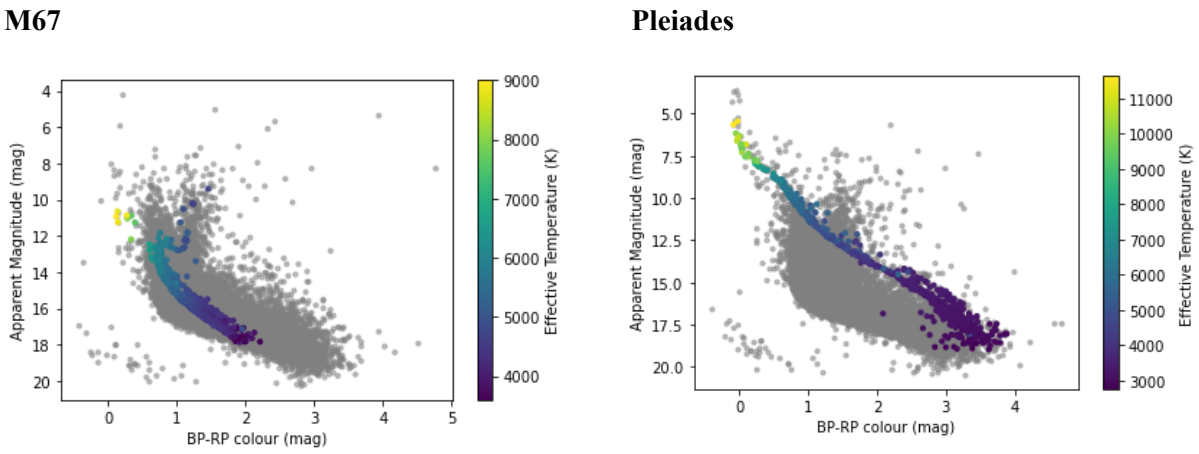


Fig.5: *Colour-Magnitude diagrams for both M67 and Pleiades open clusters. The grey dots represent the respective original subsets of data that the cluster has come from.*

Figure 5 shows a Colour-Magnitude diagram of the two clusters, after filtering for proper motions and parallax. This shows the relationship between the Absolute Magnitude and BP-RP colours. Apparent magnitude is a term referring to the perceived brightness of a star, measured in magnitude. BP-RP is the colour of each star. More specifically, how blue it is. The more blue a star is, the higher its temperature. When looking at the effective temperature of the stars, you can see that in general when the BP-RP colour decreases the temperature of the star rises. The distinct shapes of the plots that come from these clusters mean that it is a good point of reference when comparing the performance of a clustering algorithm to the true labels.

6. Detailed Analysis

6.1. Process

For the model I applied a custom process, which involved selecting the variables for the model, standardising and scaling them. Firstly, I selected the features to use for the model (pmra, pmdec and parallax). From the data cuts in the EDA, you can clearly see the difference between the cluster members and non-cluster members using these features, which shows that it will be an important determinant in cluster candidacy. The model was tested with the Right Ascension and Declination included, but the model performed better without these variables so it was left out of the model.

The data features had different distributions from each other, so having them standardised would help with balancing the effects that these features have on the clustering selection.

DBSCAN was used because it can detect outliers when clustering, which is important in this scenario since some subsets include nearly 100,000 records that try to find a ~750-size cluster. DBSCAN is a density-based clustering algorithm, which is able to filter out large amounts of ‘noisy’ data by finding closely-related data points.

The DBSCAN algorithm has two parameters, ‘minPts’ and ‘eps’. ‘Eps’ is a measure of distance to find neighbouring points, and ‘minPts’ are the minimum number of neighbouring points required to be considered in a ‘dense’ cluster. These two parameters needed to be tuned in order to perform well.

I used the F1 score of the predictions as the evaluation metric for the model. The F1 score is a balance between precision and recall. In this case, the precision is how many of the actual cluster candidates are classified correctly, and the recall is how many of the predicted cluster candidates are actual cluster candidates. This is relevant to the data since the F1 score is useful in unbalanced datasets, and each subset often has under 5% of the subset as cluster members.

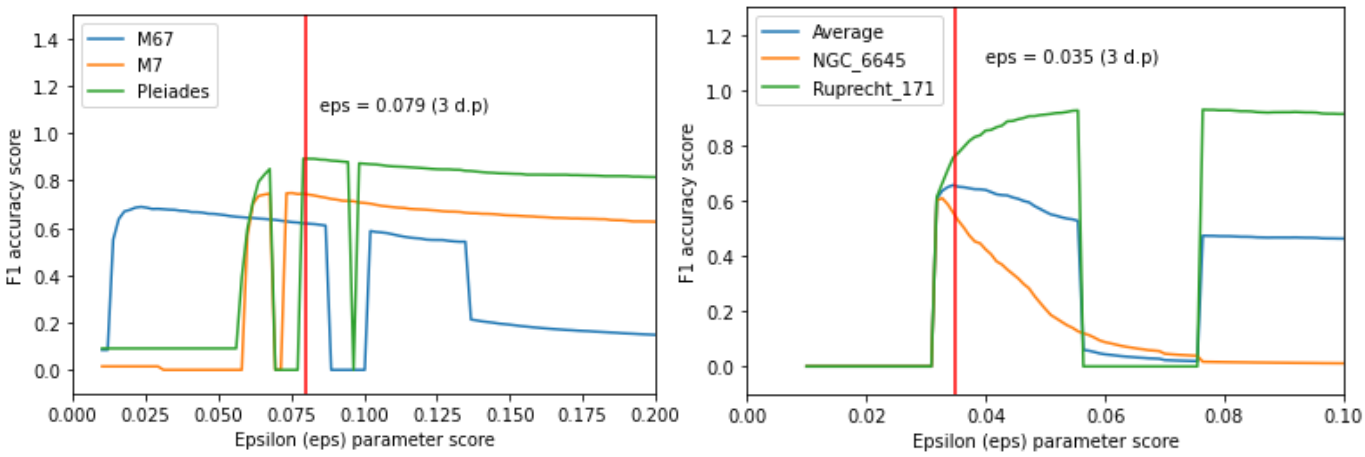
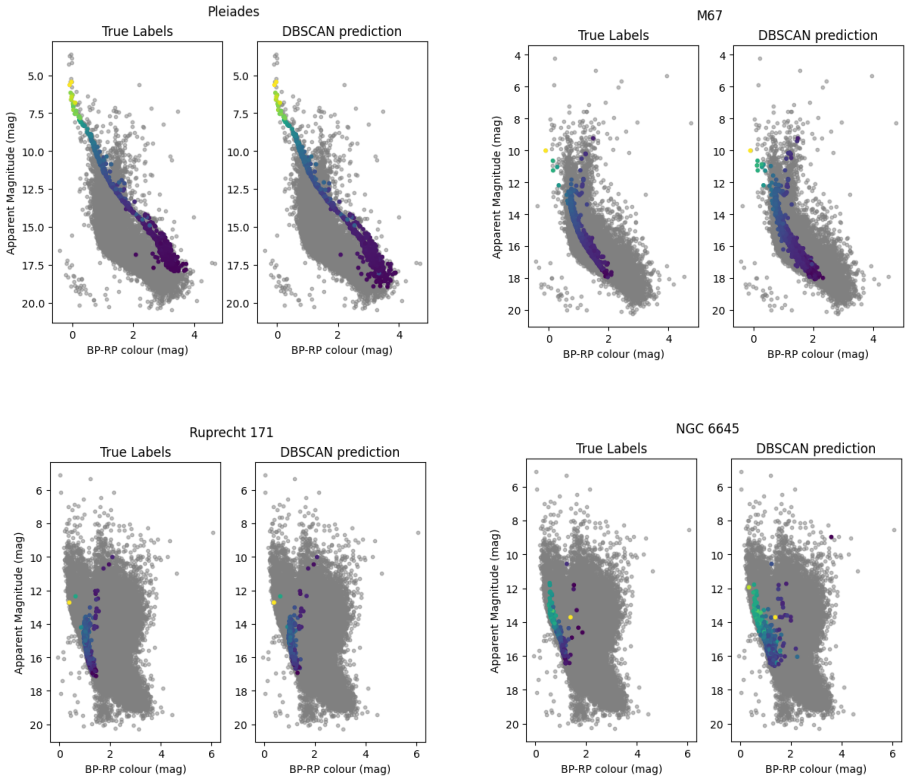


Fig.6: *Left:* DBSCAN performance for the three single-cluster subsets, with varying Epsilon values. *Right:* DBSCAN performance for the single two-cluster subset, with varying Epsilon values. The red lines highlight the optimum value of the range of parameter values tested.

The epsilon parameter value for the DBSCAN was tuned in order to get the best-performing model for the three single-cluster datasets. This value was taken from the average of the F1 scores between the three clusters, and as shown above, the maximum is roughly around 0.079, which will be used in the model for the following analysis on the single-cluster subsets.

Separate tuning was done for the two-cluster subset, and the optimum value between the two clusters was found. The two clusters seem to diverge as the epsilon value increases, with the Ruprecht cluster improving while the NGC cluster declines in accuracy. From inspecting the cluster predictions, the Ruprecht classification seemed to be capturing more cluster members as the epsilon increased, and the NGC cluster was also capturing more stars, but not those that were members of the cluster.

6.2. Results



Cluster	F1 Score (1 d.p)	Precision (1 d.p)	Recall (1 d.p)
Pleiades	89.3 %	83.6 %	95.8 %
M67	62.0 %	45.0 %	100.0 %
M7	74.5 %	61.7 %	94.1 %
Ruprecht 171	76.7 %	92.7 %	65.4 %
NGC 6645	55.0 %	40.8 %	84.4 %

Fig.7: *Top:* Colour-Magnitude diagrams for the three single-cluster subsets, comparing the true labels to the DBSCAN prediction *Bottom:* Colour-Magnitude diagram for the two-cluster subset, comparing the true labels to the DBSCAN prediction. The table on the right side displays the F1 scores for each cluster.

For the most part, the main shape of each cluster's prediction is pretty close to their true labels. The F1 scores ranged from 55% all the way to 89%, with an average score of 71.5%. The worse-performing clusters generally have a bit more thickness and are more 'bumpy' compared to the true label diagrams.

One common trend with most of the clusters is that the recall is higher than the precision. This means that most of the time, the model does well to capture the 'true' cluster members but tends to include too many that aren't 'true' cluster members. One explanation of why is that the true labels are generated on stars with magnitude 18 and lower, so the models will include some duller stars in their clustering. From this the precision will fall, and so will the F1 accuracy score. This can be seen above, where the bottom of the Colour-Magnitude diagrams show some extra predictions once the magnitude passes 18.

7 . Conclusions and Recommendations

From the Detailed Analysis we can conclude that the clustering algorithm DBSCAN is a reasonably effective method for detecting existing open clusters. We can confidently say that the proper motions and parallax are important features in open cluster membership.

Finding a model which performs well on all of the subsets of data came at a cost of individual performance. This may be improved with more research into why some performed worse than others, and it may require some feature construction or a better process (eg: ensemble learning).

A more complete true label set would improve the metric scoring of the algorithm by acquiring a sample from an astronomer, for example.

Future work may include being able to discover new clusters that are further away from us, which the DBSCAN may help in finding. Having scientific expertise would be required in order to judge if the classifications are valid, and they would be able to bring more insights about these clusters (eg: the age of the cluster).

8. Reference List

Héctor Cánovas. (2022). Use Cases. European Space Agency.
https://www.cosmos.esa.int/web/gaiausers/archive/use-cases

René Andrae. (2022). GSP-Phot (Generalized Stellar Parametrizer). Max Planck Institute.
https://www.mpia.de/gaia/projects/gsp

Cantat-Gaudin, T., Vallenari, A., Sordo, R., et al. (2018). Characterising open clusters in the solar neighbourhood with the Tycho-Gaia Astrometric Solution.
https://ui.adsabs.harvard.edu/abs/2018A%26A...615A..49C/abstract

Krone-Martins, A., & Moitinho, A. (2014). UPMASK: unsupervised photometric membership assignment in stellar clusters. https://ui.adsabs.harvard.edu/abs/2014A%26A...561A..57K/abstract

Cantat-Gaudin, T., & Anders, F. (2020). Clusters and mirages: cataloguing stellar aggregates in the Milky Way. https://ui.adsabs.harvard.edu/abs/2020A%26A...633A..99C/abstract

9 . Appendices

Appendix A - Code examples to extract datasets (Pleiades as an example)

Go to https://gea.esac.esa.int/archive/ and navigate to **Search > Advanced(ADQL)** and paste the following:

```
SELECT source_id, ra, dec, pmra, pmdec, parallax, phot_g_mean_mag, bp_rp, teff_gspphot
FROM gaiadr3.gaia_source
WHERE
CONTAINS(
        POINT('ICRS',gaiadr3.gaia_source.ra,gaiadr3.gaia_source.dec),
        CIRCLE(
                'ICRS',

COORD1(EPOCH_PROP_POS(56.601,24.114,7.3640,19.9970,-45.5480,6.5700,2000,2016.0)),

COORD2(EPOCH_PROP_POS(56.601,24.114,7.3640,19.9970,-45.5480,6.5700,2000,2016.0)),
                2)
)=1
AND abs(pmra_error/pmra)<0.10
AND abs(pmdec_error/pmdec)<0.10
AND abs(parallax_error/parallax)<0.10
```

This gets you the raw data around the centre of Pleiades, with 2 degrees of range and filtering of missing data and data with high error rates.

To generate the labels, go to http://tapvizier.cds.unistra.fr/adql/?%20J/A+A/633/A99/members and paste in the following:

```
-- output format : csv
SELECT "J/A+A/633/A99/members".Source
FROM "J/A+A/633/A99/members"
WHERE "J/A+A/633/A99/members".Cluster = 'Melotte_22'
AND Proba > 0.6
```

This will get the IDs of the stars with membership probability of the Pleiades cluster above 60%.