**Classification of Star Cluster Members using Gaia DR3**

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

Table of Contents

[Executive Summary 2](#_Toc148535206)

[Background 3](#_Toc148535207)

[Data Description 3](#_Toc148535208)

[Ethics and Privacy 5](#_Toc148535209)

[Exploratory Data Analysis (EDA) 6](#_Toc148535210)

[Detailed Analysis Results 9](#_Toc148535211)

[Conclusions and Recommendations 13](#_Toc148535212)

[Reference List 13](#_Toc148535213)

[Appendices 14](#_Toc148535214)

# Executive Summary

Leveraging the Gaia DR3 data, the aim of this research is to accurately classify stars as belonging to a specific star cluster within a specified region of the sky. The Gaia DR3 data is publicly available online through the European Space Agency (ESA). For this research, 6877 instances and 12 features were extracted from the full Gaia DR3 dataset.

The 12 features extracted from the data include one unique identified *‘source\_id’*, and 11 astronomical features relating to the position, motion, colour, and temperature of stars. The target column, specifying whether a star is a cluster member or not, was not extracted from GAIA DR3, but from a secondary source (Parkosidis, 2022) by merging the two sources on a unique identifier, *‘source\_id’*.

These 6877 instances constitute two separate datasets. 3575 of these instances are stars in or near the NGC 7789 star cluster. The remaining 3302 instances are stars in or near the Trumpler 5 star cluster. Both these datasets are comprised of the same 12 features. Two separate classification models were trained, one on each dataset, in the aims of classifying a star as being a part of the NGC 7789 cluster or Trumpler 5 cluster, in their respective regions of space.

In both datasets (NGC 7789 and Trumpler 5), one of the astronomical features, *‘radial\_velocity’* (radial velocity) were almost entirely missing. 81% of the radial velocity values were missing for stars in the NGC 7789 region, and 93% were missing for stars in the Trumpler 5 region. As a result, radial velocity was not used a feature to train either of the models. Among the other features, there was no more than 10% of data missing for either of the datasets. These missing data were imputed using median values. In addition, 264 instances were removed from the NGC 7789 data, as they had erroneous right ascension *(‘ra’)* values. All numerical variables, in both datasets, were then standardised.

For each classification model, 60% of the data was used as a training set, 20% as a validation set, and 20% as a final test set. Extreme gradient boosting (XGBoost) was the algorithm used for both models. Synthetic sampling was used to train the Trumpler 5 model, as the data suffered from severe class imbalance. AUC-ROC wasused to assess the NGC 7789 model as it did not suffer from such a heavy class imbalance, whilst PR-AUC was used to assess the Trumpler 5 model. Since the two models themselves are not being compared (as they are used for different regions), the use of different metrics to assess their performance is reasonable.

The NGC 7789 model achieved an AUC-ROC of 0.8890 and was balanced in terms of the incorrect predictions it made. On the other hand, the Trumpler 5 model achieved a PR-AUC of 0.7152 but suffered from bias. It failed to correctly classify a handful of actual cluster members as cluster members, which is due to the class imbalance in the dataset. Overall, the two models had good performance, and will be useful in classifying stars as cluster members or not.

|  |  |  |  |
| --- | --- | --- | --- |
| Star Cluster Region | AUC-ROC | False Positives | False Negatives |
| NGC 7789 | 0.8890 | 55 | 49 |
| Star Cluster Region | PR-AUC | False Positives | False Negatives |
| Trumpler 5 | 0.7152 | 29 | 83 |

*Table 1: NGC 7789 model and Trumpler 5 model performance/results.*

# Background

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.

**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.

# Data Description

This research has made use of data from the European Space Agency (ESA) mission Gaia ([https://www.cosmos.esa.int/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](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.

In the Gaia DR3 data, there are over 150 different features available. Only eleven features were selected with an additional column that was a unique source identifier. 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. For this research, 6877 instances and 12 features were extracted from the larger Gaia DR3 data using SQL queries in Python. 3575 of these instances are stars in or near the NGC 7789 star cluster, and 3302 instances are stars in or near the Trumpler 5 star cluster.

The Gaia DR3 data does not contain star cluster labels, so this extraction was done by extracting source identifiers from a secondary dataset containing star cluster labels and merging with the Gaia data on *‘source\_id’*. This secondary dataset was created by Adam Parkosidis from the University of Amsterdam (<https://github.com/adamparkosidis/statistical-analysis-Gaia-data/tree/main> ).

This secondary dataset contained 401448 instances, containing astronomical information (across 24 features) extracted from the Gaia data. There was column called *‘Cluster’*, containing the cluster the star belonged to, and a column *‘PMemb’*, describing the membership probability. Instances that were in the NGC 7789 and Trumpler 5 clusters were extracted. Out of these instances, stars with a *‘PMemb’* value less than 1.0 were classed as non-members. *‘PMemb’* was numerical discrete, ranging from 0.0 to 1.0, increasing from 0.0 in 0.1 increments.

After merging, two separate datasets were created, each containing data about stars in or near each of the two cluster regions. Each dataset contains information about individual stars in the same region of space. *‘source\_id’* was consequently dropped as a feature from each dataset, as it contains no astronomical information useful for model training.

*Table 2* lists the features and their descriptions that constituted the labelled datasets. The two datasets had the same features.

|  |  |
| --- | --- |
| 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). G band measures the brightness of objects in the green part of the electromagnetic spectrum. |
| phot\_bp\_mean\_mag | Mean magnitude in the BP band. BP band measures the brightness of objects in the blue part of the electromagnetic spectrum. |
| phot\_rp\_mean\_mag | Mean magnitude in the RP band. RP band measures the brightness of objects in the blue part of the electromagnetic spectrum. |
| 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 2: Features in each labelled dataset used for classification model training.*

Each feature in *Table 2* is numerical, except *‘Cluster’*, which is nominal categorical. 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 3: Class balance in each dataset.*

*‘Table 3’* illustrates the class imbalance in each of the datasets. The Trumpler 5 region data is somewhat imbalanced, potentially 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 4: Features with missing data (and how many missing instances), for each cluster dataset.*

*Table 4* highlights that *‘radial\_velocity’* across both is afflicted with lots of missing data. Considering that most of the data is missing for that column, this feature was dropped and not used for model training. The remaining features had a significantly smaller proportion of missing values. Therefore, the missing values for each feature were imputed with the median value of that feature.

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 ranges between 358 and 360 degrees. The data was then consequently filtered to exclude the 264 erroneous instances.

# 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.

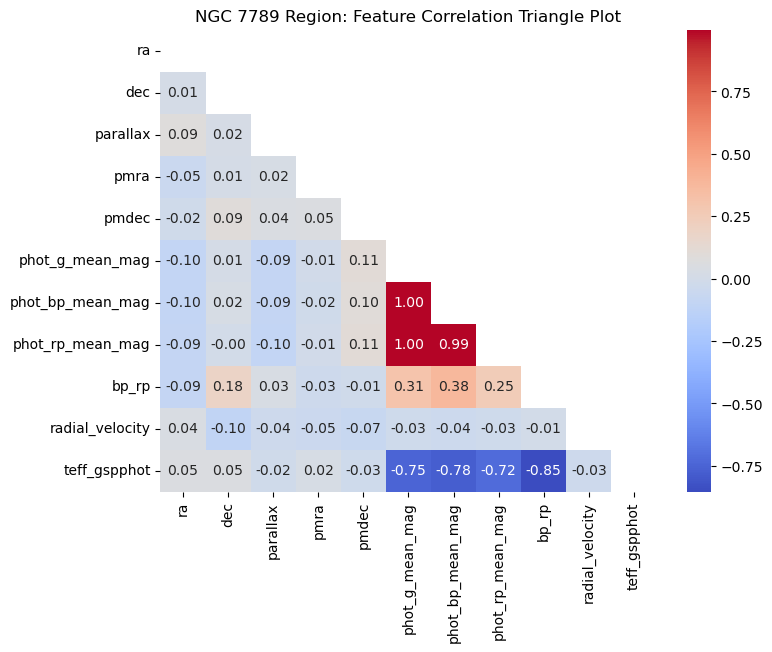
Like ethical concerns, the usual privacy concerns relating to most types of data do not necessarily apply to the Gaia DR3 data. An important thing to consider is that although the data is not personal, it is important to ensure that other sources indirectly connected to individuals or organisations are 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 by storing data and results on local and remote servers. Access to these servers is limited and is only available to individuals directly working on this research.

# Exploratory Data Analysis (EDA)

**Feature Correlations**

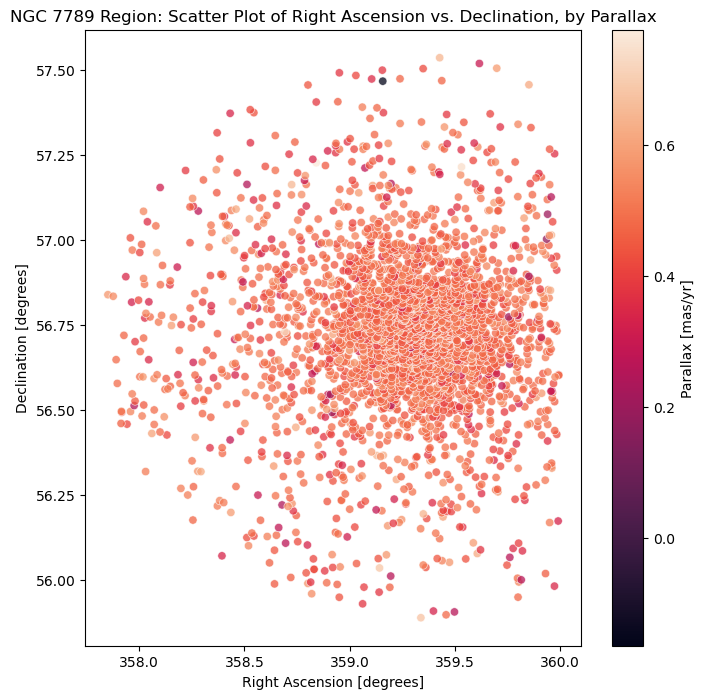
Among both cluster regions, the highest correlations were between the three features that provide information about the observed brightness of a star. *‘phot\_g\_mean\_mag’*, *‘phot\_bp\_mean\_mag’*, and *‘phot\_rp\_mean\_mag’* each had extremely strong, positive correlations with one another, ranging 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. A similar correlation matrix for the Trumpler 5 data revealed similar patterns, hence it was not included.

Notable correlations unique to the NGC 7789 data were medium to strong negative correlations between *‘teff\_gspphot’* and the four features mentioned previously. The correlation is expected considering that *‘teff\_gspphot’* is estimated from *‘bp\_rp’*.

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

**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.

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

The lack of extreme values in these plot indicate, at least in the case of these specific star clusters, that parallax combined with right ascension and declination are not extremely useful predictors of cluster membership.

*Figure 2: NGC 7789 cluster region visualised.*

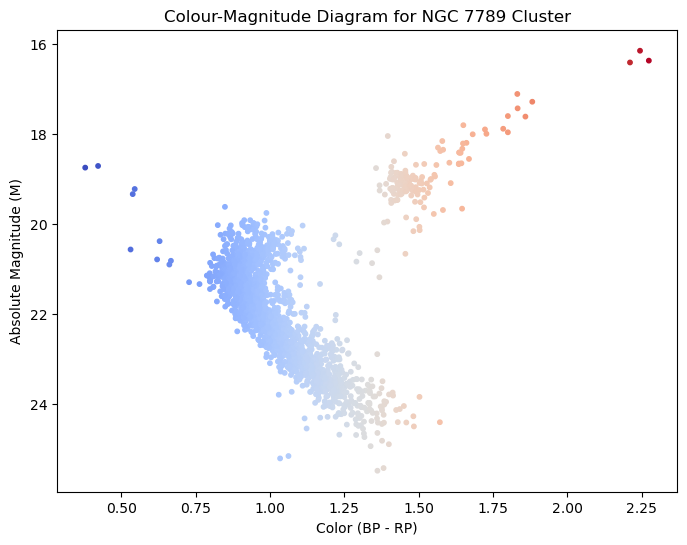
**Colour-magnitude diagrams (CMD)**

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.

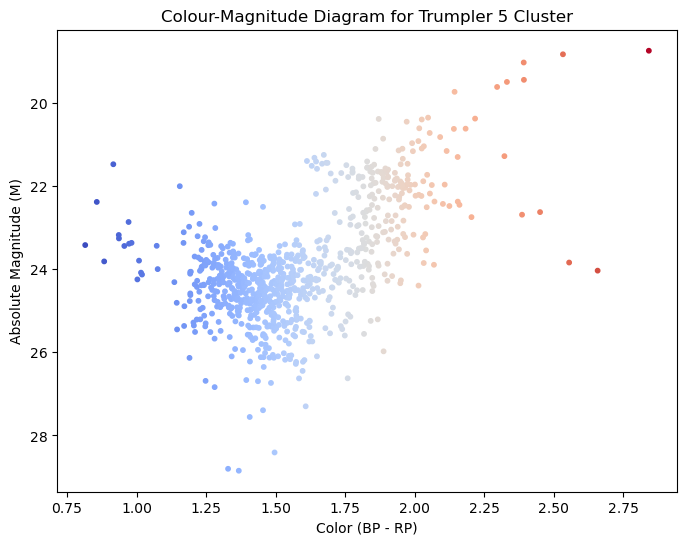
The ‘main sequence’ in a CMD is the prominent diagonal feature that extends from the top left (bright, blue stars) to the bottom right (dim, red stars). The main sequence consists of stars that are in the prime of their lives and are actively fusing hydrogen into helium in their cores.

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).

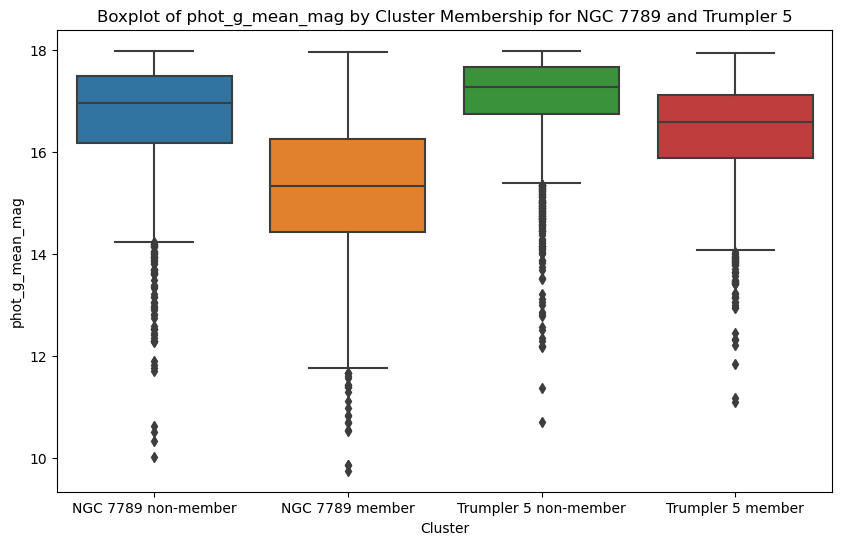
*Figure 3* illustrates a clear turn-off point for stars in the NGC 7789 cluster region, along with some blue ‘stragglers’ just to the left and above the main sequence. There is notable a gap in space between stars on the main sequence and stars turning off the main sequence. Initially, this was assumed 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 (Kiss, 2016; Mermilliod, 2000).

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

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

*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 an older cluster, meaning many stars have already transitioned to the later stages of their evolution. Consequently, the brightest and bluest stars are less numerous, resulting in the less defined main sequence.

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



# 

*Figure 5: Difference in ‘phot\_g\_mean\_mag’ between members and non-members, for each cluster region.*

*Figure 5* illustrates that cluster members, for both cluster regions, generally tend to have a lower *‘phot\_g\_mean\_mag’* value than non-cluster members. There is minimal overlap between members and non-members in the NGC 7789 data. The median *‘phot\_g\_mean\_mag’* value for NGC 7789 members is considerably lower than the median for NGC 7789 non-members. The difference between Trumpler 5 members and non-members is not as distinct, but still carries the same trend that the NGC 7789 data does. This indicates that *‘phot\_g\_mean\_mag’* could be a useful predictor of cluster membership.

# Detailed Analysis Results

Extreme gradient boosting models (XGBoost) were the choice of algorithm for this task. XGBoost naturally can handle common issues found in datasets, such as feature multicollinearity, class imbalance, and missing values. Although the algorithm can work with these issues, steps were taken beforehand to remedy these problems, such as data preprocessing and feature selection.

Each dataset was split into train, validation, and test sets with a 60-20-20 split. *Table 5* shows the number of instances in each set, for each cluster region dataset.

*Table 5: Train-validation-test splits for each cluster region dataset.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset  (Star Cluster Region) | Train Set | Validation Set | Test Set | Total |
| NGC 7789 | 2118 | 530 | 663 | 3311 |
| Trumpler 5 | 2112 | 529 | 661 | 3302 |

*‘radial\_velocity’* was dropped because of excessive missing data. For the other features, missing values were imputed using the median value of that feature and all features were standardised (after imputation).

AUC-ROC was used to assess the NGC 7789 model, whilst AUC-PR was used to assess the Trumpler 5 model. This is since the Trumpler 5 model was working with imbalanced data. The values calculated are a result of 5-fold stratified cross-validation with the mean value across the folds being the overall value. The uncertainty of these values is provided as the standard deviation across the folds.

For each of the two datasets, three initial models, without hyperparameter tuning, were evaluated on the validation set. These models were each trained on a different subset of features (informed by EDA).

* Model 1 was trained on all features.
* Model 2 was trained on the same features as Model 1, but also excluded *‘phot\_bp\_mean\_mag’, ‘phot\_rp\_mean\_mag’, ‘bp\_rp’, and ‘teff\_gspphot’.*
* Model 3 was trained on the same features as Model 2 but retained *‘teff\_gspphot’.*

**NGC 7789 Cluster Region Classification Model**

As discovered in the EDA, 264 instances in the NGC 7789 region data contained erroneous right ascension values. These instances were removed from the data.

*Table 6: AUC-ROC scores for the three models, with uncertainty.*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Number of Features | Validation AUC-ROC | Validation AUC-ROC Uncertainty |
| Model 1 | 10 | 0.9178 | 0.0168 |
| Model 2 | 6 | 0.9137 | 0.0150 |
| Model 3 | 7 | 0.9165 | 0.0175 |

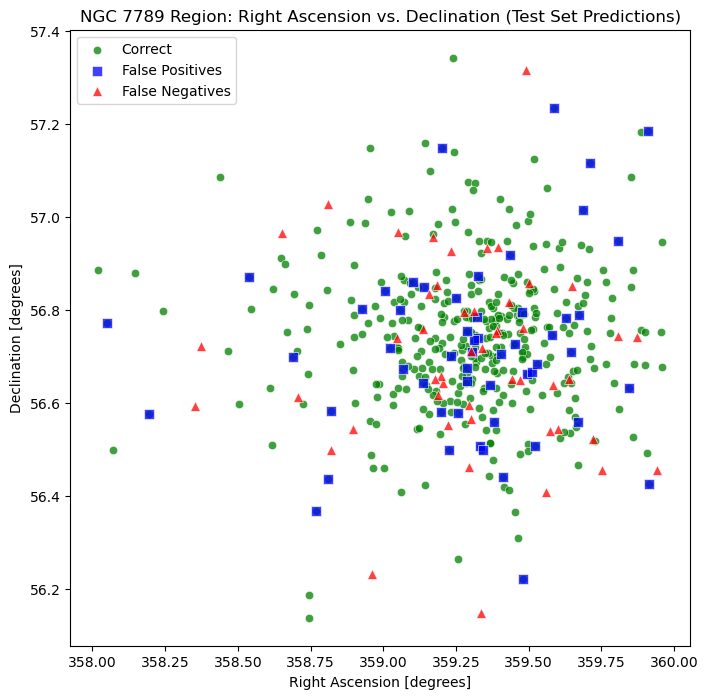
The three models all had extremely similar performance, as seen in *Table 6*. Since the difference in performance is so small between the models, it makes sense to choose the simplest model (least features), as simpler models reduce the risk of overfitting. As a result, Model 2 was developed further with hyper-parameter tuning.

GridSearchCV was used to tune the number of estimators, maximum depth, and learning rate hyper-parameters in XGBoost. The best tuned model used 100 estimators, with a maximum depth of 3, and a learning rate of 0.1.

*Table 7: AUC-ROC scores for the tuned model, on validation and test sets.*

|  |  |  |  |
| --- | --- | --- | --- |
| Validation AUC-ROC | Validation AUC-ROC Uncertainty | Test AUC-ROC | Test AUC-ROC Uncertainty |
| 0.9235 | 0.0112 | 0.8890 | 0.0321 |

The tuned model had an improved performance on the validation set. As seen in *Table 7*, the performance very slightly dropped when evaluated on the test set, however, the drop is not large enough to indicate overfitting.

*Figure 6* depicts the predictions made by the tuned model. There is an even spread of where the erroneous predictions in terms of their right ascension and declination positions. In total there were 49 false negatives and 55 false positives.

This indicates that there seems to be no significant bias in the model, as there seems to be no area in the region where incorrect predictions are made more frequently, nor does it incorrectly classify positive or negative instances more frequently than the other.

Extracting feature importance values for this model revealed that *‘phot\_g\_mean\_mag’* was considered the most useful predictor by a considerable margin. The feature importance scores of this model are presented in Appendix A.

*Figure 6: Visualisation of test set predictions.*

**Trumpler 5 Cluster Region Classification Model**

*Table 8: PR-AUC scores for the three models, using synthetic sampling.*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Synthetic Sampling | Validation PR-AUC | Validation PR-AUC Uncertainty |
| Model 1 | False | 0.6704 | 0.1010 |
| Model 1 | True | 0.7086 | 0.0695 |
| Model 2 | False | 0.6205 | 0.0597 |
| Model 2 | True | 0.6430 | 0.0635 |
| Model 3 | False | 0.6517 | 0.0538 |
| Model 3 | True | 0.6700 | 0.0503 |

Synthetic sampling was implemented using SMOTE (Synthetic Minority Over-sampling Technique) and Random Under Sampling. Together, these techniques allowed for a balanced dataset (same number of members and non-members) to be created. This was achieved by first oversampling the minority class to have 80% the number of instances in the majority class. Then, the majority class was under sampled to match the number of instances in the minority class. The final dataset had 3957 instances overall (655 more instances than the original data).

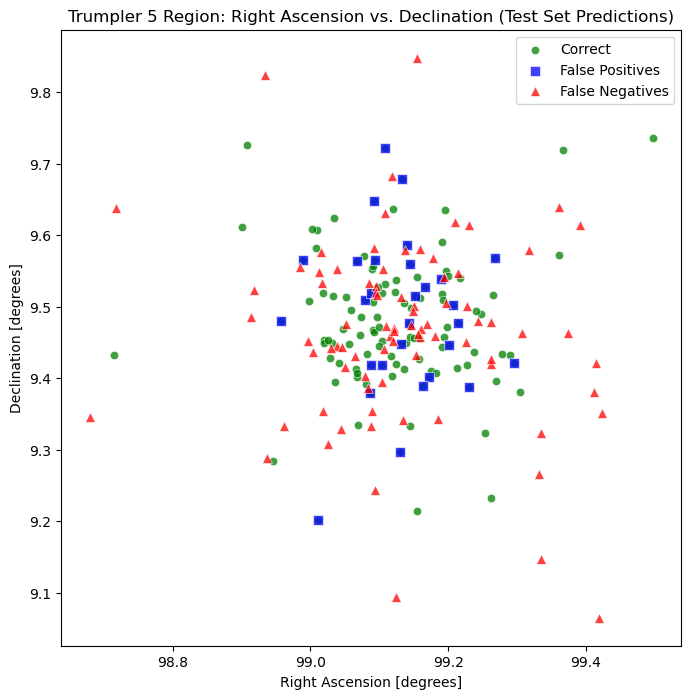
*Table 8* highlights that the utilisation of synthetic sampling boosted the performance of each model. Model 1, the most complex model (trained on all features) had the best performance, followed by Model 3 – both using synthetic sampling. Model 3 is a simpler model but had a PR-AUC value 0.3086 lower than that of Model 1. Therefore, even though Model 3 was simpler, Model 1 was developed further using hyper-parameter tuning due to its increased performance.

Similar to the NGC 7789 model, the number of estimators, maximum depth, and learning rate parameters were tuned. The best tuned model used 100 estimators, with a maximum depth of 3, and a learning rate of 0.1.

*Table 9: PR-AUC scores for the tuned model, on validation and test sets.*

|  |  |  |  |
| --- | --- | --- | --- |
| Validation PR-AUC | Validation PR-AUC Uncertainty | Test PR-AUC | Test PR-AUC Uncertainty |
| 0.6809 | 0.0671 | 0.7152 | 0.0662 |

The tuned model had a slightly poorer performance on the validation set. However, the performance on the test set was the best performance the model had produced so far, as seen in *Table 9*. This is good sign, as it indicates that the model generalises well to unseen data.

*Figure 6* depicts the predictions made by the tuned model. Most of the false positives are found within the centre of the cluster, whilst there is a wider spread in the positions of false negatives. There were 83 false negatives, compared to only 29 false positives. This is consistent with the reasonably high PR-AUC score, which quantifies the accuracy of the model’s positive predictions.

There seems to be bias present in the model, as it fails to classify a considerable number of member stars as members. This is most likely a result of the class imbalance in the data, which even remedied with synthetic sampling, did not fully resolve the issue.

*Figure 6: Visualisation of test set predictions.*

Extracting feature importance values for this model revealed that *‘phot\_g\_mean\_mag’* was considered the most useful predictor, followed closely by *‘pmra’* and *‘pmdec’*. The feature importance scores of this model are presented in Appendix B.

# 

# Conclusions and Recommendations

The performance of these models indicate that it is possible to classify stars as cluster members or not in certain regions of space, to a reasonably high degree. The accuracy of which this can be done depends on the data available for a certain region, and each region would require a separate model with unique parameter, features, and data.

The NGC 7789 classification model had a stronger performance compared to the Trumpler 5 classification model. This is since the data used to train the Trumpler 5 model was severely imbalanced, and as a result the model failed to accurately classify several actual member stars as being a member. Based on the performance of the Trumpler 5 model, it would be worth further investigating class imbalance remediation techniques specific to astronomical data.

The most useful predictor in both of the models was *‘phot\_g\_mean\_mag’*. The next 5 most useful features were also the same: *‘pmdec’, ‘pmra’, ‘parallax’, ‘dec’, ‘ra’.* These features relate to the motion, position, and brightness of a star. Stars in the same cluster tend share similar motions, positions, and brightness. The order of importance of these features differs between the models. Interestingly, these 6 features were the features used to train the NGC 7789 model, whilst the Trumpler 5 model was trained on all 10 features, as it had poorer performance when trained on just those 6 features.

The limitations of the results produced in this report arise mainly from data quality. The class imbalance in the Trumpler 5 data somewhat affected the results of the final Trumpler 5 model. The absence of radial velocity as a feature for both models due to large amounts of missing data is unfortunate. Radial velocity could potentially be useful in improving classification accuracy; however, this feature has a lot of missing data, not just in the data used in this project, but in the entire Gaia DR3 data.

In conclusion, this research demonstrates the potential for accurately classifying stars as cluster members or not of particular clusters in specific regions of space. The NGC 7789 model showcased strong performance, while the Trumpler 5 model faced challenges due to class imbalance. Further investigation into class imbalance remediation techniques for astronomical data is warranted. Despite some limitations, this study exemplifies the promise of machine learning in understanding and identifying celestial objects in the universe.

# Reference List

European Space Agency (ESA). (2023). Gaia Data Release 2. [Data set]. ESA. <https://www.cosmos.esa.int/web/gaia/dr3>

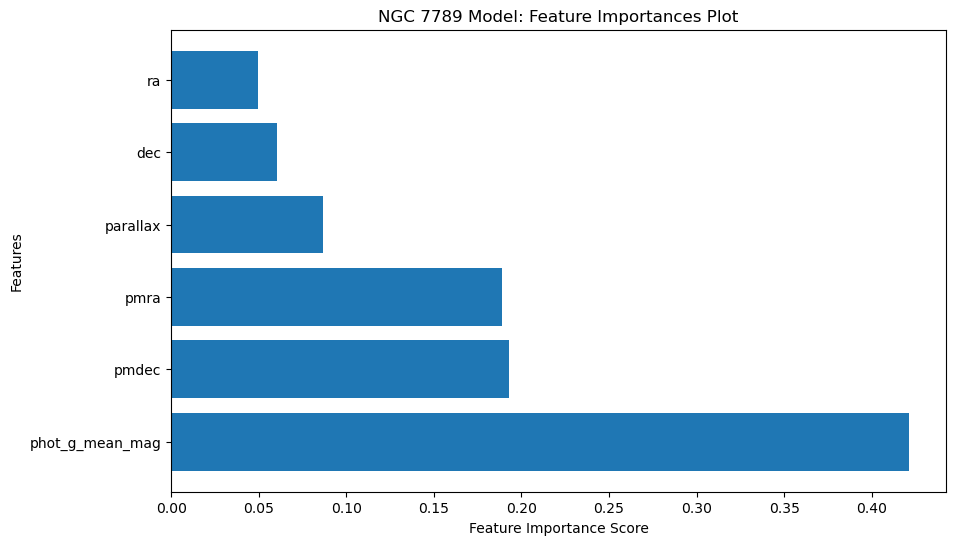
Kiss, P. (2016, January 31). *NGC 7789.*   
<https://deepeye.hu/en/drawings/ngc7789.html>

Mermilliod, J-C. (2000). On the peculiar red clump morphology in the open clusters NGC 752 and NGC 7789. *Astronomy and Astrophysics.* <https://www.researchgate.net/publication/2219669_On_the_peculiar_red_clump_morphology_in_the_open_clusters_NGC_752_and_NGC_7789>

Parkosidis, A. (2022). *An Analysis of Gaia’s data, estimating the distance of cluster NGC2506 using Bayesian statistics*. [Data set]. GitHub. <https://github.com/adamparkosidis/statistical-analysis-Gaia-data/tree/main>

# Appendices

**Appendix A**Feature importance scores for NGC 7789 cluster region model.



**Appendix B**Feature importance scores for Trumpler 5 cluster region model.

