

Capstone (Undergrad Thesis)

---



**MINERVA**<sup>®</sup>  
UNIVERSITY

# A Semi-supervised Machine Learning Approach for Open Star Cluster Member Determination

Md Mahmudunnobe

Major: CS (Data Science) / NS (Atoms and Molecules)



**MINERVA**<sup>®</sup>  
UNIVERSITY

## Background: Open Cluster

---

A group of star, which

- Born from the same molecular clouds
  - Similar Age, Similar Structure
- Loosely Held together by Mutual Gravity
  - Similar Position in sky
  - Similar distance
  - Similar velocity



# Background: ML models

---



**MINERVA**<sup>®</sup>  
UNIVERSITY

## 1. Unsupervised Learning

- No Labelled Training Data
- Gaussian Mixture Model (GMM)  
(Gao, 2020; Agarwal et al. 2020)
- DBSCAN algorithm  
(Gao 2014; Bhattacharya et al. 2017)
- UPMASK  
(Cantat-Gaudin et al. 2018, 2020)

## 2. Supervised Learning

- Labelled Training Data
- Random Forest  
(Gao, 2018)
- Artificial Neural Network (ANN)  
(Castro-Ginard et al. 2018)



**MINERVA<sup>®</sup>**  
UNIVERSITY

# What: Hypothesis

---

- Past works mainly used unsupervised models  
(Agarwal et al. 2020, Cantat-Gaudin et al, 2020, etc.)
- A supervised model is more accurate and easier to validate  
(Jain et al, 1999, Reddy et al, 2018)
- Use of an additional supervised model increased the members  
(Gao, 2018b; Castro-Ginard et al. 2018, Mahmudunnobe et al, 2021)
- **Hypothesis:** Based on these evidences, I argue that a combination of unsupervised and a supervised model would be a better approach to detect members of open star clusters.

# What: Specific Focus

---



**MINERVA**<sup>®</sup>  
UNIVERSITY

## Goal of the Project:

- Develop a ***workflow*** to combine unsupervised and supervised models
- Develop or Suggest ***metrics*** to compare between different models
- Suggest the ***best practices*** for applying a specific model

(Chosen Model to Explore: GMM, Random Forest)

## Why: Why it matters now?

---



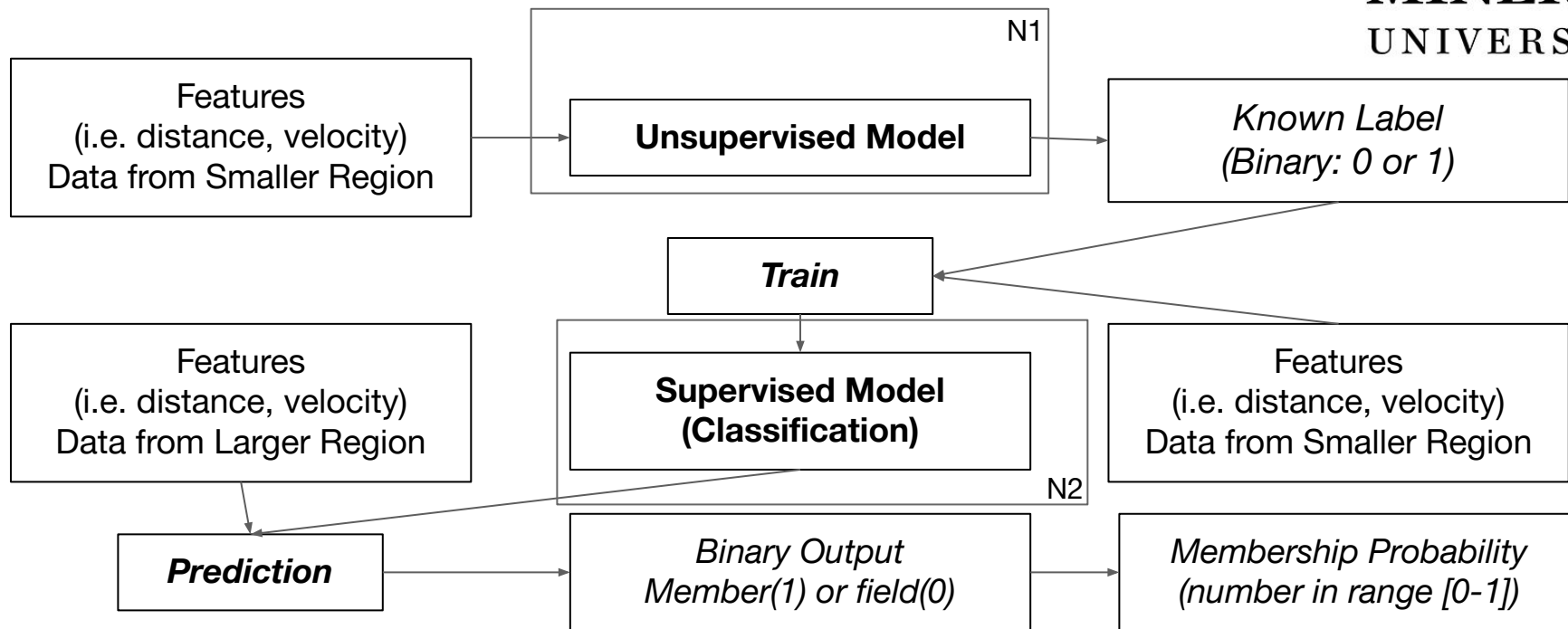
**MINERVA**<sup>®</sup>  
UNIVERSITY

- GAIA provides the precise data for distances, positions, velocities of stars
- Having confirmed members allow further studies:
  - Understanding Star Formation
  - Understanding Galaxy Formation
  - Modeling Stellar Evolution



**MINERVA<sup>®</sup>**  
UNIVERSITY

## How: Workflow



# How: Metrics

---



**MINERVA**<sup>®</sup>  
UNIVERSITY

## Unsupervised Model

- Members should be more compact (smaller SD)
- Modified Silhouette Score (MSS)

$$\frac{1}{K} \sum_{i=1}^K \frac{(SD_{i,field} - SD_{i,member})}{\max(SD_{i,field}, SD_{i,member})}$$

## Supervised Model

- Want to avoid false positive
- Precision

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

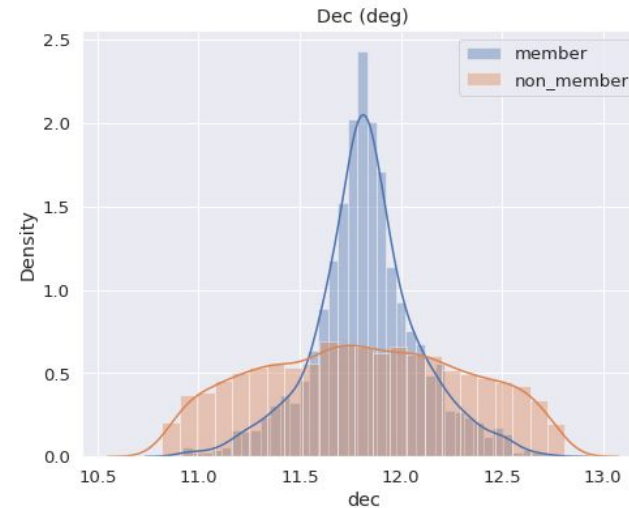
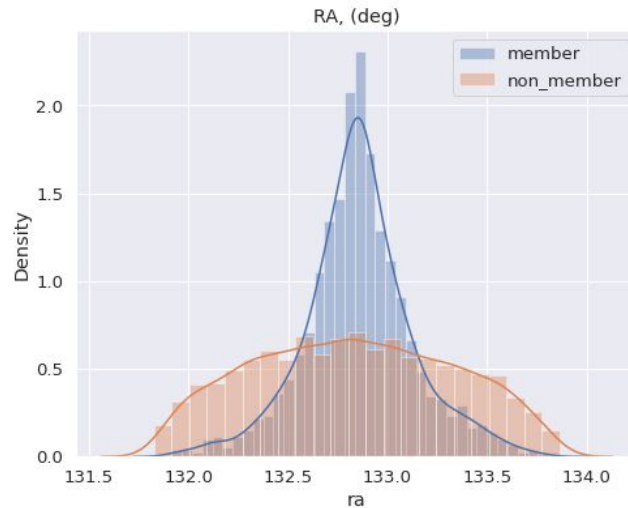


# GMM: Modified Replication



**MINERVA**  
UNIVERSITY

1. **Feature Selection:** Remove ra and dec from the features
  - As they overlap for a small search radius

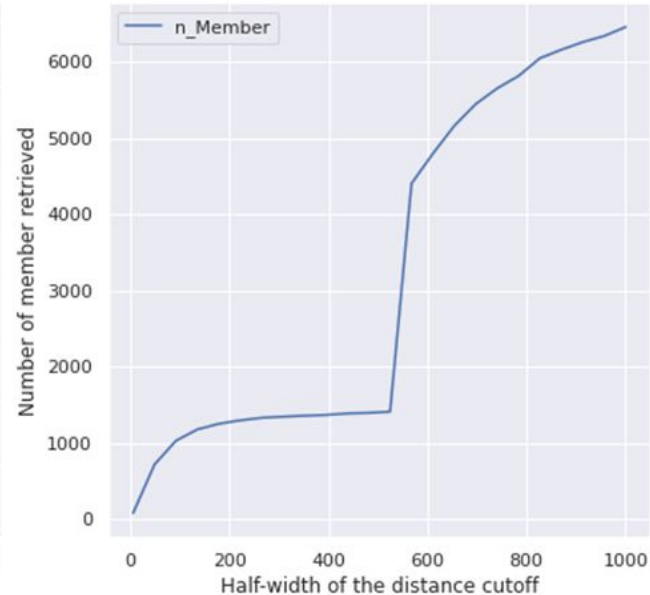
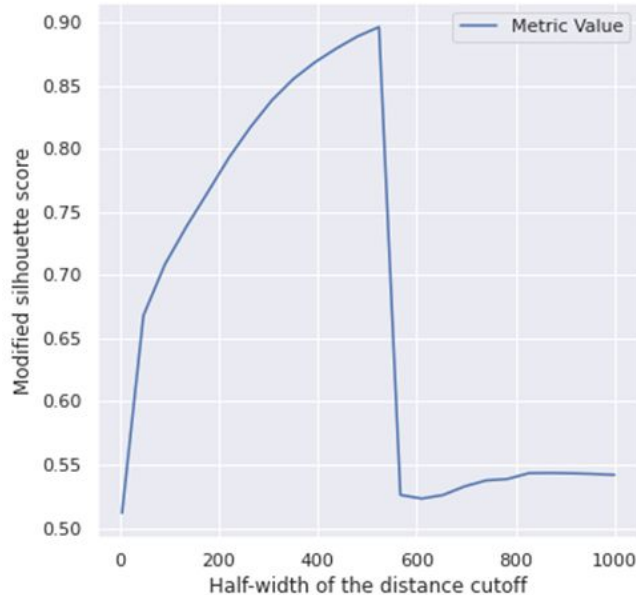


# GMM: Modified Replication



**MINERVA**  
UNIVERSITY

2. Choose the **distance cutoff** using MSS and number of members

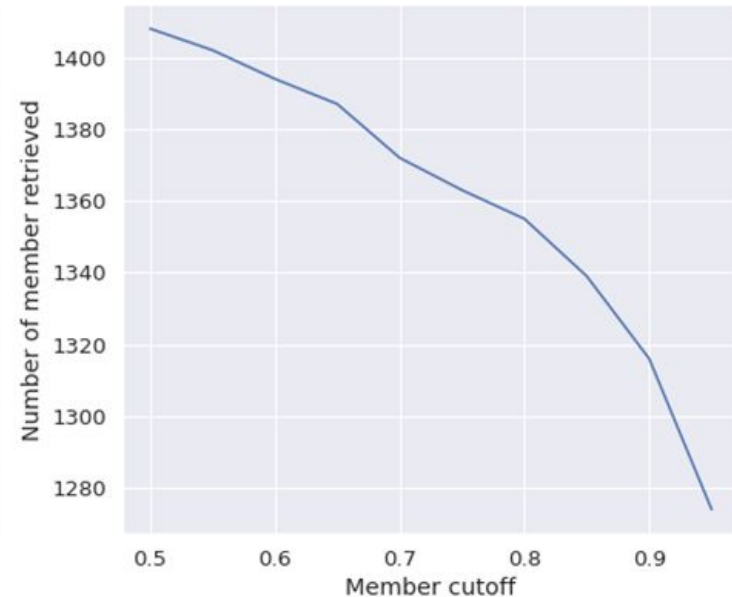
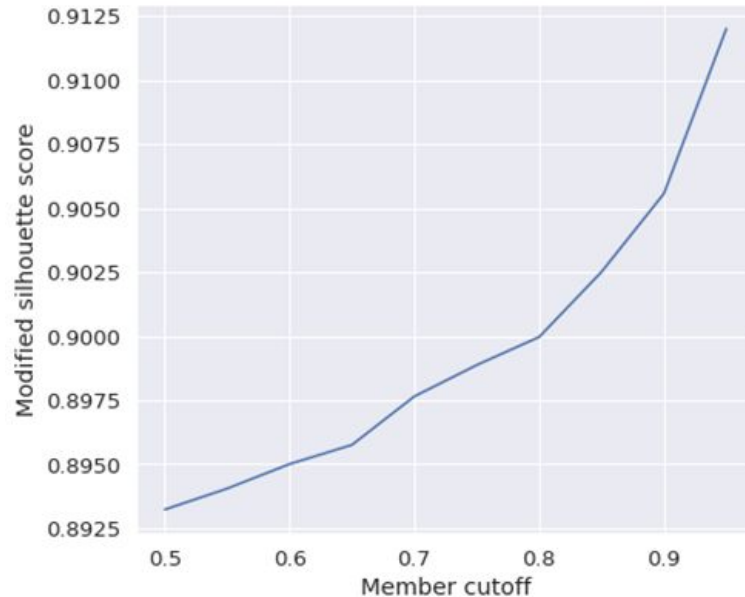


# GMM: Modified Replication



**MINERVA**<sup>®</sup>  
UNIVERSITY

3. Choose the **member threshold** using MSS and number of members

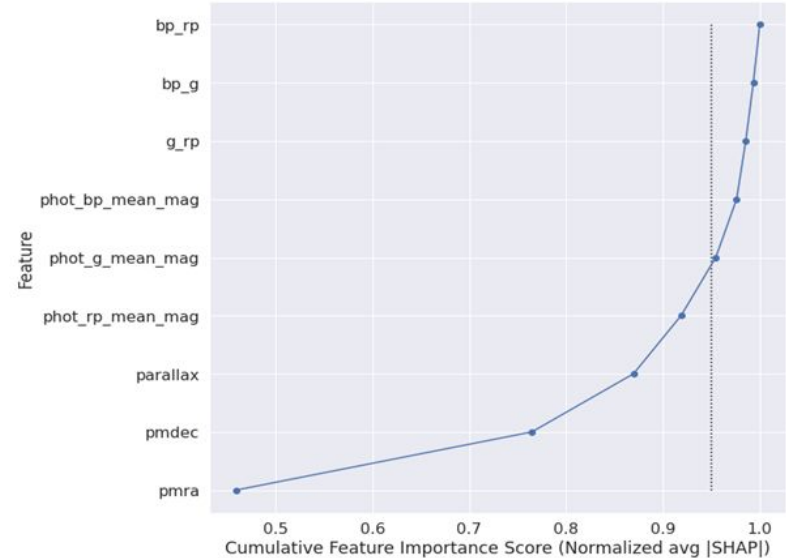
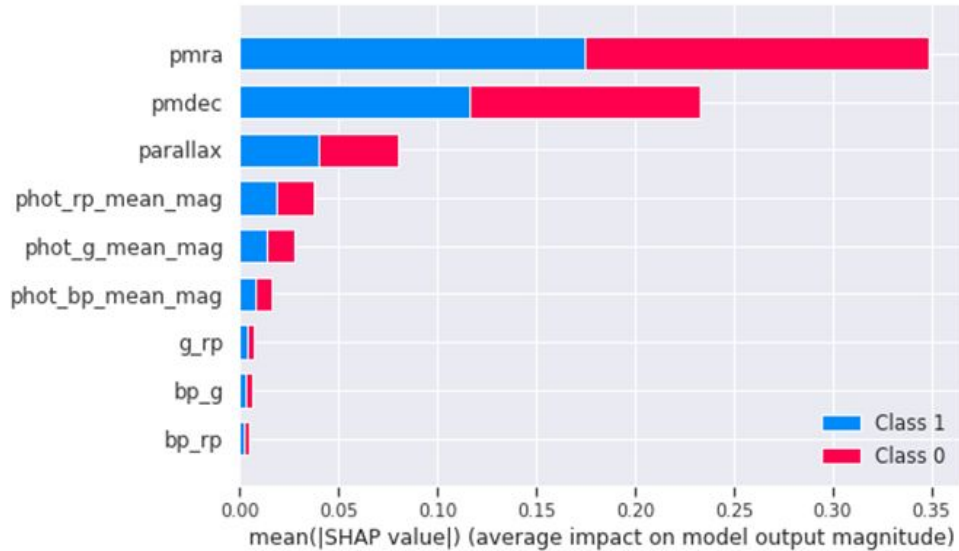


# RF: Modified Replication



**MINERVA<sup>®</sup>**  
UNIVERSITY

## 1. Feature Selection: SHAP feature importance





**MINERVA**<sup>®</sup>  
UNIVERSITY

## RF: Modified Replication

---

### 2. Divide into test and training data

- for unbiased performance evaluation
- Stratified split for maintaining class distribution

Proportion of class:

original target data:

1 0.501605

0 0.498395

test\_targets:

1 0.501538

0 0.498462

train\_targets:

1 0.501633

0 0.498367



## RF: Modified Replication

### 3. Optimize Hyperparameter using Cross Validation

<i>N_estimators</i> (Number of decision trees)	500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000
<i>max_features</i> (Number of features in each tree)	'sqrt', 1, 2, 3, 4, ..., n_feature
<i>max_depth</i> (Maximum depth of the tree)	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None
<i>min_samples_leaf</i> (Minimum samples required for a leaf node)	1, 2, 4
<i>min_samples_split</i> (Minimum samples required to split a node)	2, 5, 10
<i>bootstrap</i> (Whether bootstrapping the samples)	True, False

```
{'n_estimators': 500,  
'min_samples_split': 5,  
'min_samples_leaf': 4,  
'max_features': 'sqrt',  
'max_depth': 20,  
'bootstrap': True}
```



## RF: Modified Replication

### 3. Optimize Hyperparameter using Cross Validation

<i>N_estimators</i> (Number of decision trees)	500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000
<i>max_features</i> (Number of features in each tree)	'sqrt', 1, 2, 3, 4, ..., n_feature
<i>max_depth</i> (Maximum depth of the tree)	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None
<i>min_samples_leaf</i> (Minimum samples required for a leaf node)	1, 2, 4
<i>min_samples_split</i> (Minimum samples required to split a node)	2, 5, 10
<i>bootstrap</i> (Whether bootstrapping the samples)	True, False

```
{'n_estimators': 500,  
'min_samples_split': 5,  
'min_samples_leaf': 4,  
'max_features': 'sqrt',  
'max_depth': 20,  
'bootstrap': True}
```

# Result



**MINERVA<sup>®</sup>**  
UNIVERSITY

	Direct Replication M67	Modified Replication M67	NGC 3766
MSS (GMM) @ 0.6	0.70	0.89	0.61
MSS (GMM) @ 0.95	0.73	0.91	0.78
Precision (RF) in train data	1.00	1.00	1.00
Precision (RF) in test data	-	1.00	1.00
Number of Members	1375 (0.8)	1447 (0.8)	6352 (1)

Number of Members	M67	NGC 3766
Cantat-Gaudin paper (2020)	845	1368

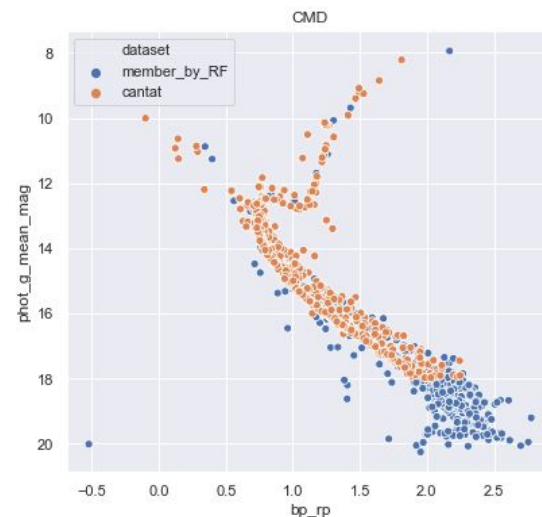
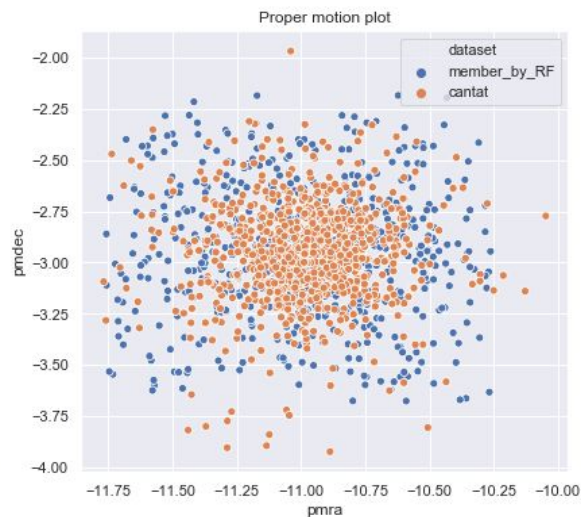
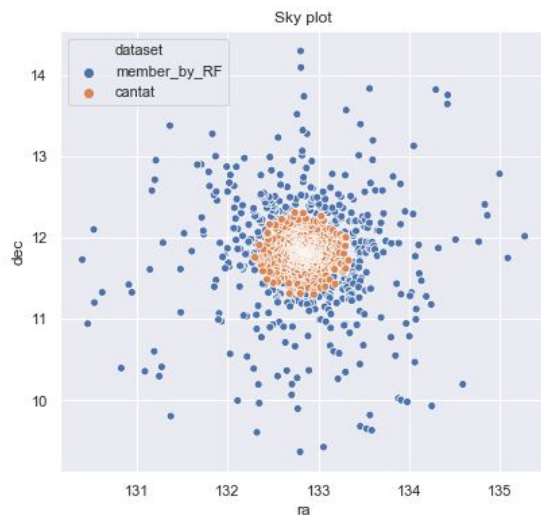


# Result: Cantat Benchmark



**MINERVA<sup>®</sup>**  
UNIVERSITY

Distribution Predicted Members and Cantat Members of M67



# Result: Cantat Benchmark



**MINERVA<sup>®</sup>**  
UNIVERSITY

Distribution Predicted Members and Cantat Members of NGC 3766





**MINERVA<sup>®</sup>**  
UNIVERSITY

## Conclusion

---

- More members can be found at the outer sky region and fainter magnitudes
- Need to compare between a number of model using these metrics to find the optimal model
- For unsupervised model, MSS can be used for feature selection and hyperparameter optimization.
- For supervised model, SHAP feature importance for feature selection and cross-validation for optimizing parameter.