Capstone Oral Defense

MINERVA®
UNIVERSITY

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

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

What: Hypothesis

MINERVA®
UNIVERSITY

- 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®
UNIVERSITY

Goal of the Project:

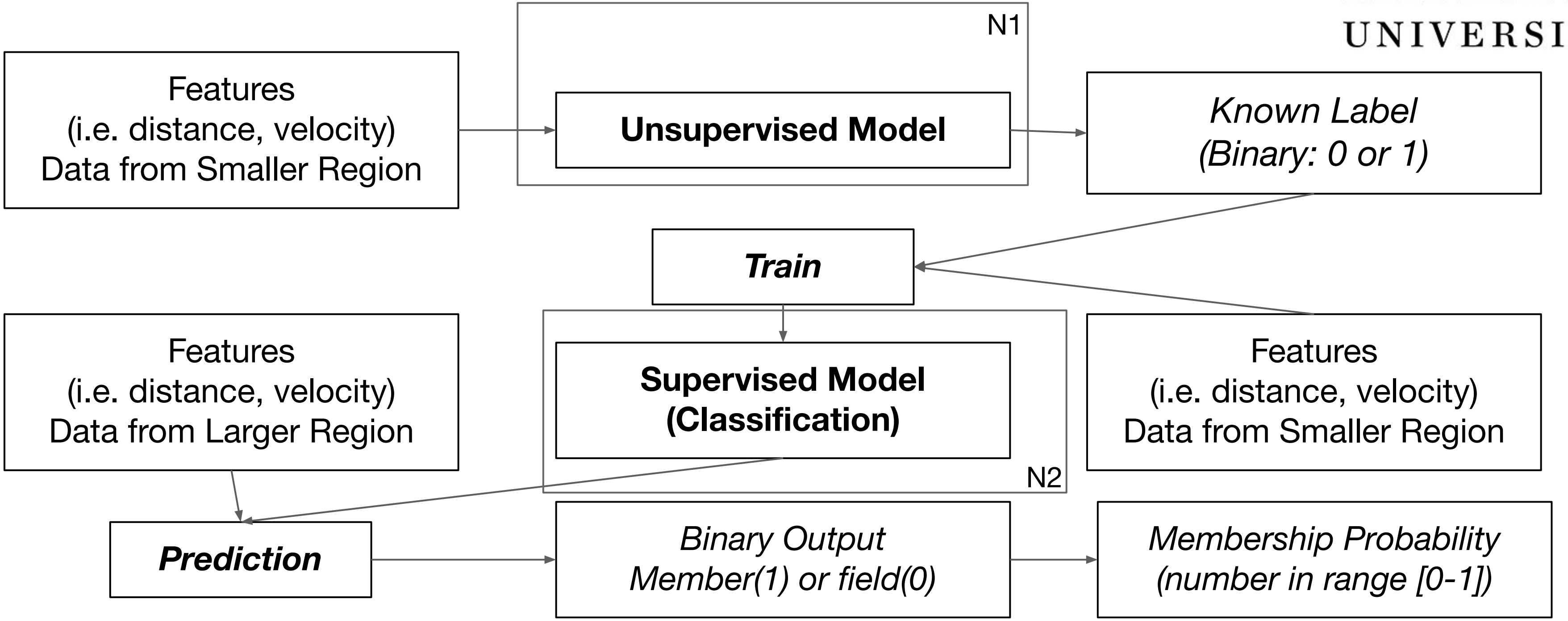
- Develop a ***workflow*** to combine unsupervised and supervised models
- Develop or Suggest ***metrics*** to compare between different models
- Suggest the ***best practices*** for applying a specific model

  (Chosen Model to Explore: GMM, Random Forest)

Why: Why it matters now?

- 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

How: Workflow



Features
(i.e. distance, velocity)
Data from Smaller Region

N1

Unsupervised Model

Known Label
(Binary: 0 or 1)

Train

Features
(i.e. distance, velocity)
Data from Larger Region

Supervised Model
(Classification)

N2

Features
(i.e. distance, velocity)
Data from Smaller Region

Prediction

Binary Output
Member(1) or field(0)

Membership Probability
(number in range [0-1])

How: Metrics

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

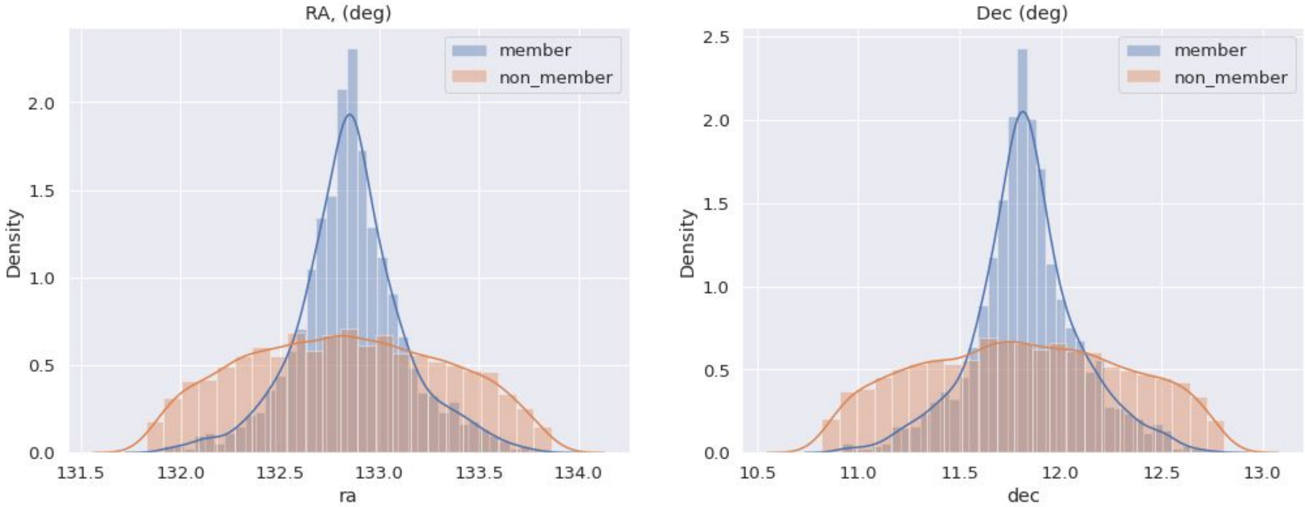
- Want to avoid false positive
- Precision

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

GMM: Modified Replication

1. **Feature Selection:** Remove ra and dec from the features

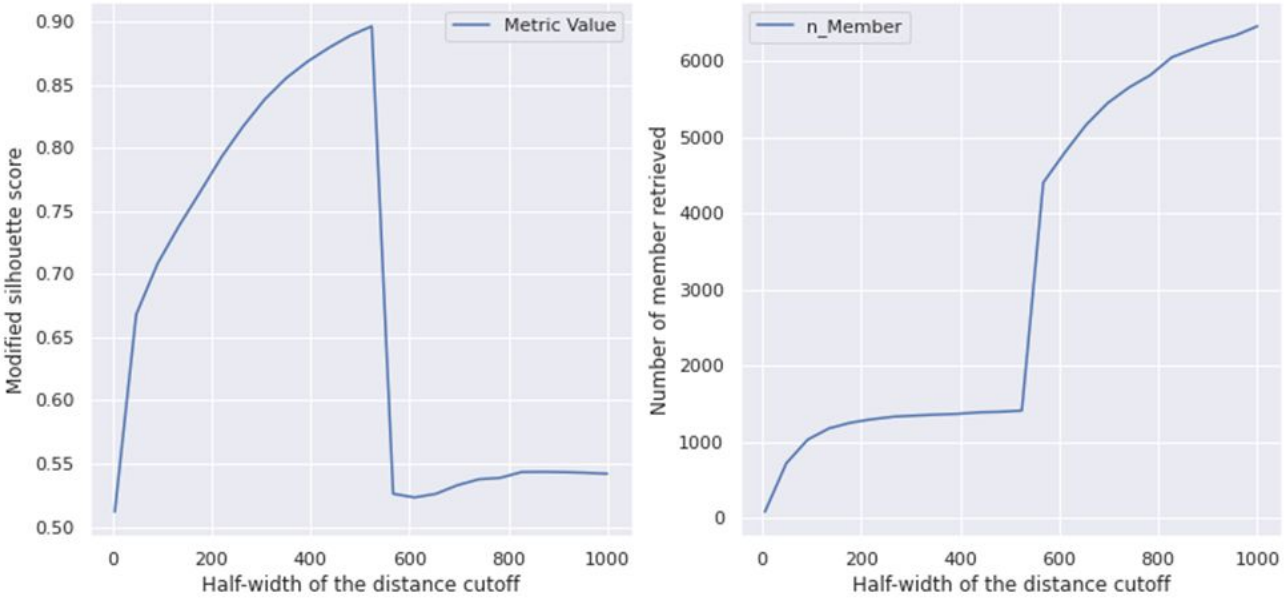
- As they overlap for a small search radius



GMM: Modified Replication

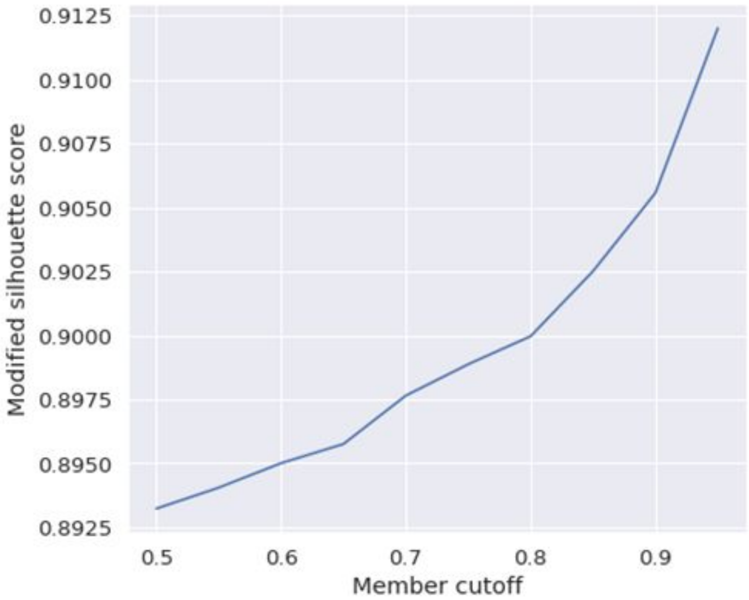
MINERVA®
UNIVERSITY

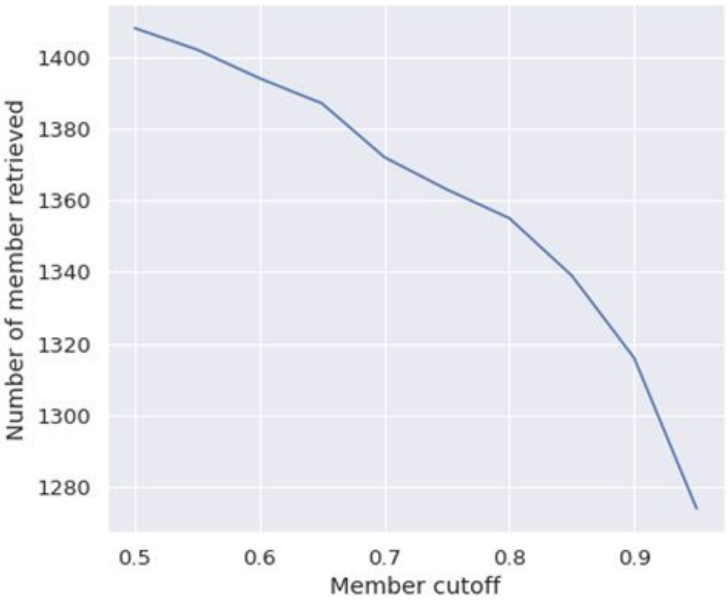
2. Choose the **distance cutoff** using MSS and number of members



GMM: Modified Replication

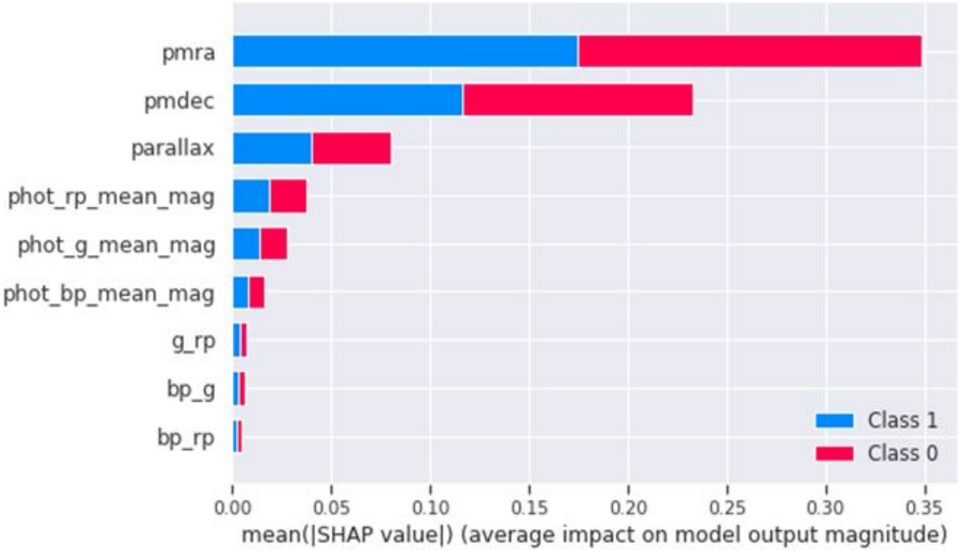
3. Choose the **member threshold** using MSS and number of members

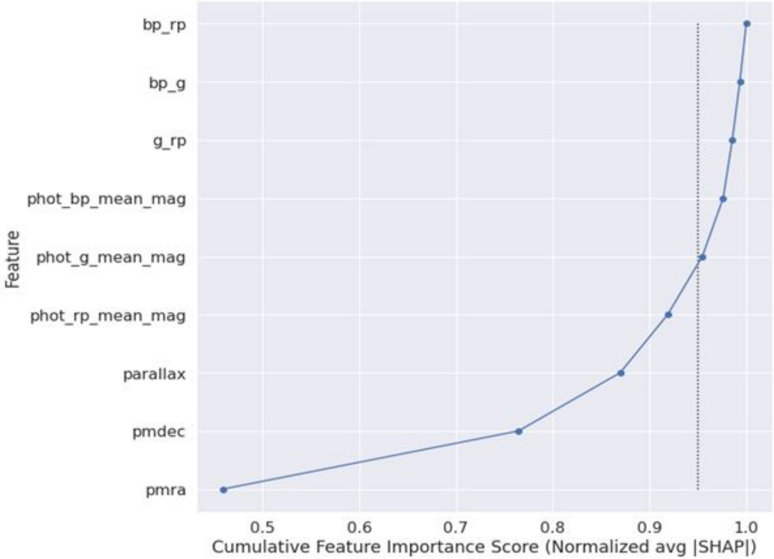




RF: Modified Replication

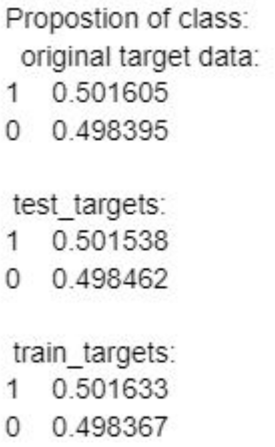
1. **Feature Selection**: SHAP feature importance





RF: Modified Replication

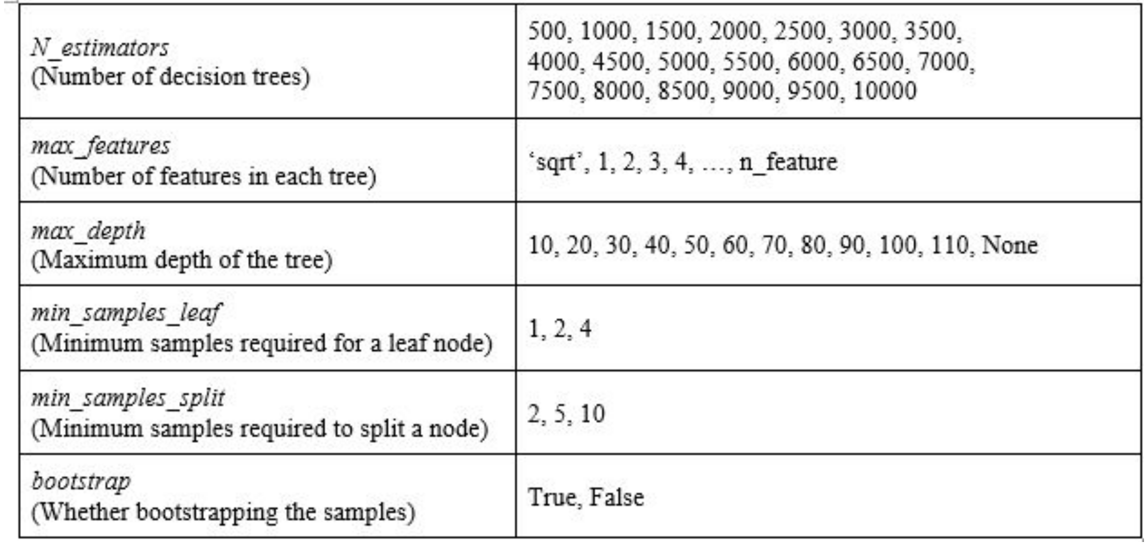
2. Divide into test and training data
   - for unbiased performance evaluation
   - Stratified split for maintaining class distribution



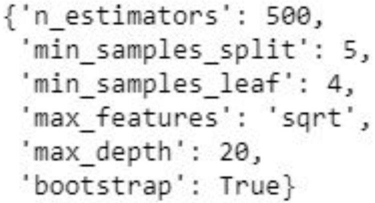
```
Propostion of class:
 original target data:
1    0.501605
0    0.498395

 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



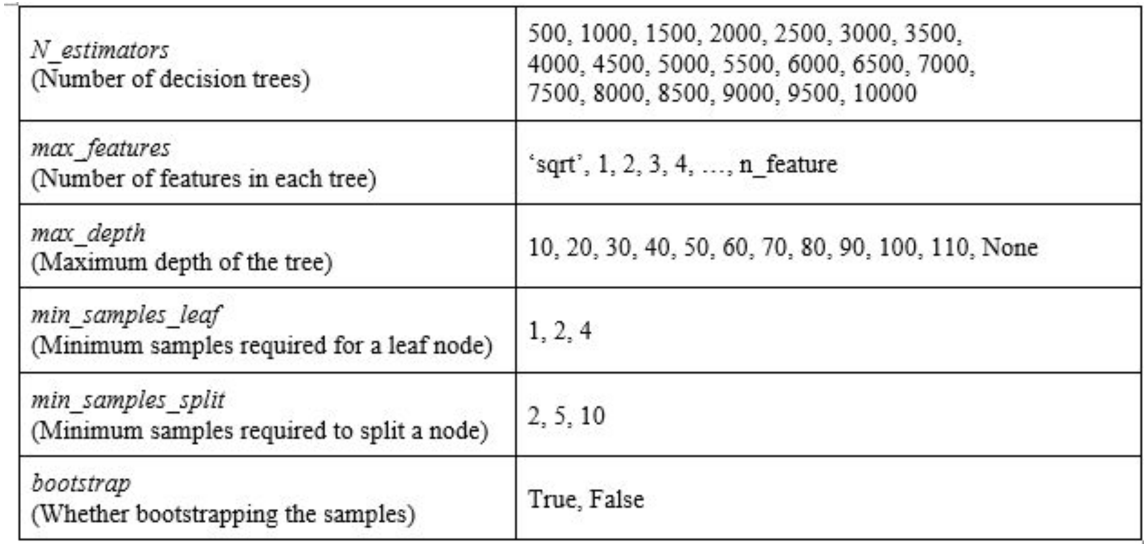
N_estimators (Number of decision trees)	500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000
max_features (Number of features in each tree)	‘sqrt’, 1, 2, 3, 4, …, n_feature
max_depth (Maximum depth of the tree)	10, 20, 30, 40, 50, 60,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



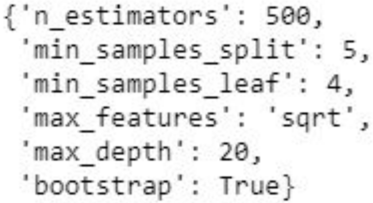
```
{'n_estimators': 500,
 'min_samples_split': 5,
 'min_samples_leaf': 4,
 'max_features': 'sqrt',
 'max_depth': 20,
 'bootstrap': True}
```

RF: Modified Replication

3. **Optimize Hyperparameter** using Cross Validation

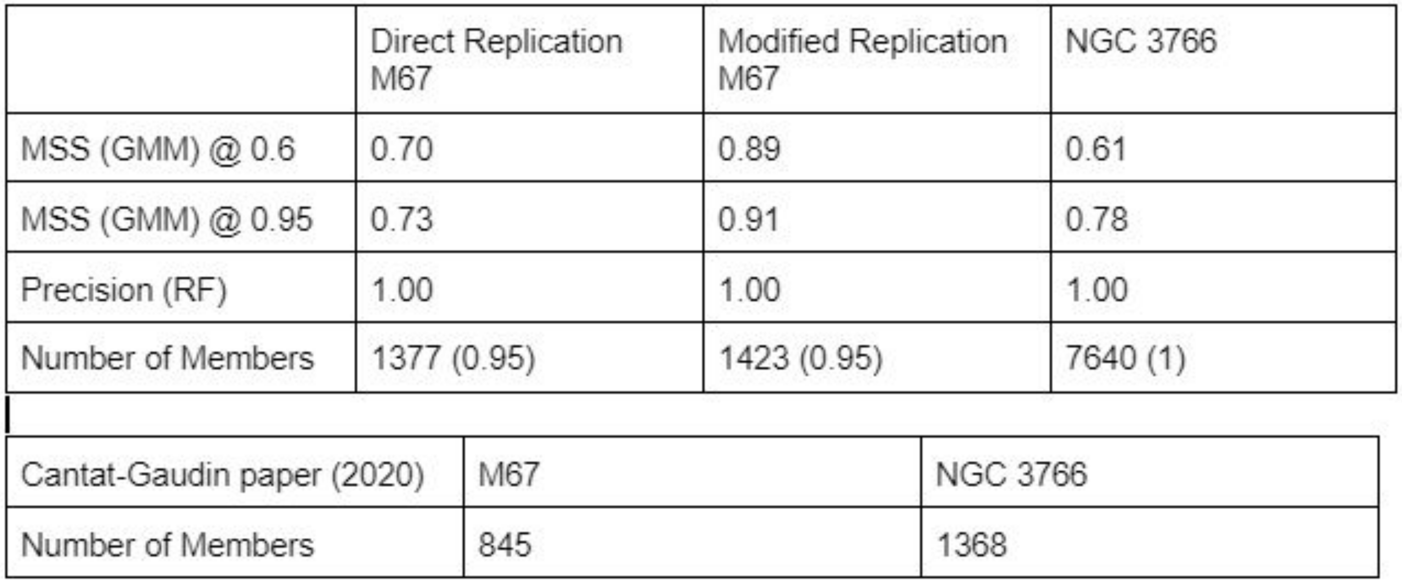


N_estimators (Number of decision trees)	500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000
max_features (Number of features in each tree)	‘sqrt’, 1, 2, 3, 4, …, n_feature
max_depth (Maximum depth of the tree)	10, 20, 30, 40, 50, 60,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

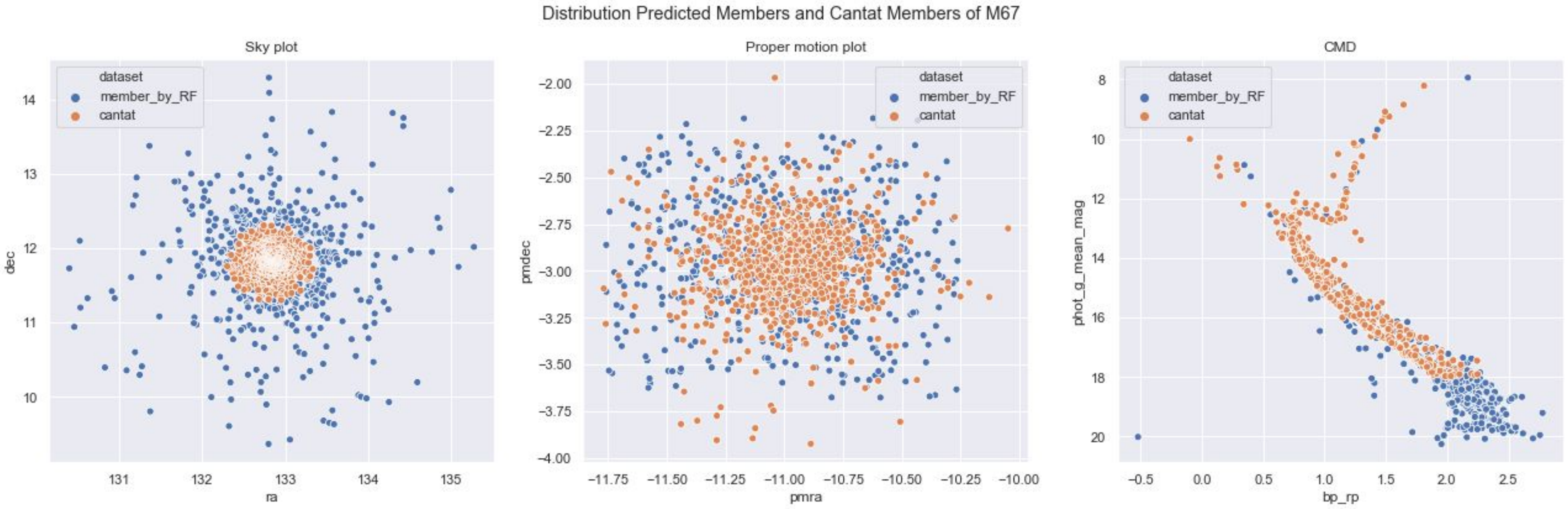
Result



	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)	1.00	1.00	1.00
Number of Members	1377 (0.95)	1423 (0.95)	7640 (1)

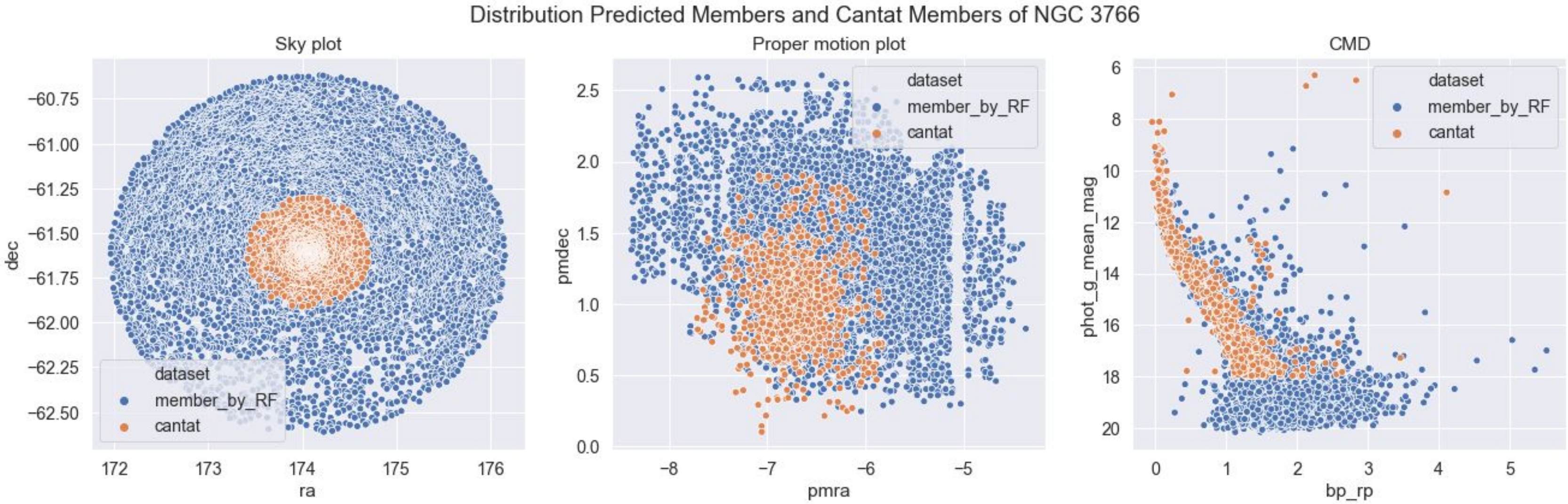
Cantat-Gaudin paper (2020)	M67	NGC 3766
Number of Members	845	1368

Result: Cantat Benchmark



Result: Cantat Benchmark

MINERVA®
UNIVERSITY



Conclusion

MINERVA®
UNIVERSITY

- 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 optimizating parameter.