GAIA's VARIABLE STARS MEET MACHINE LEARNING

L. Rossi[1], L. Santo[1], S. Vaccaro[1]

M. Brescia[1,2], M. Marconi[2], V. Ripepi[2]

[1] *Department of Physics «E. Pancini» - University of Napoli "Federico II", Via Cinthia 21, 80126, Napoli, Italy*
[2] *INAF - Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131, Napoli, Italy*

Contacts:

- lukarossi98@gmail.com
- lorenzosanto96@gmail.com
- sim.vaccaro@studenti.unina.it

ABSTRACT

The GAIA DR3 suffers from a high fraction of missing values that prevent a machine learning classification without loss of information and data.
We present an application of *Random Forest* and sk-learn *Iterative Imputer* on a GAIA DR3 sample: the aim of the work was to correctly classify *RR Lyrae* among the other variable stars and verify the possibility of using the whole dataset including objects with missing data.
The results show good performances for the classification and a minimal dataset degradation due to the imputation.

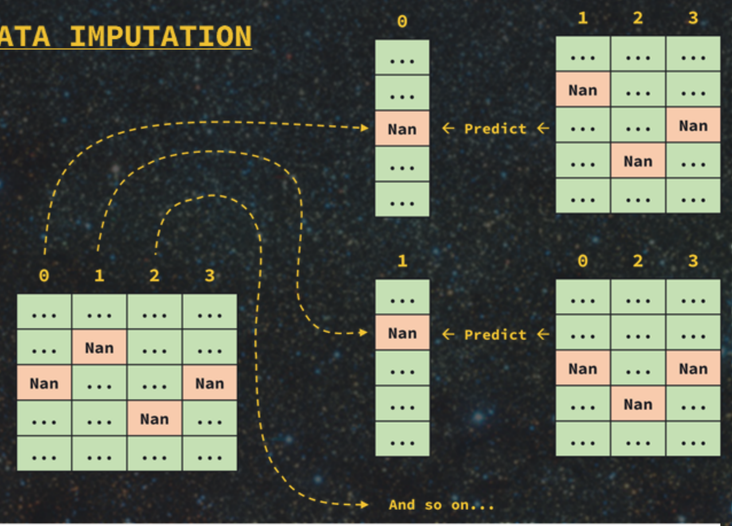
⚠ WARNING ⚠

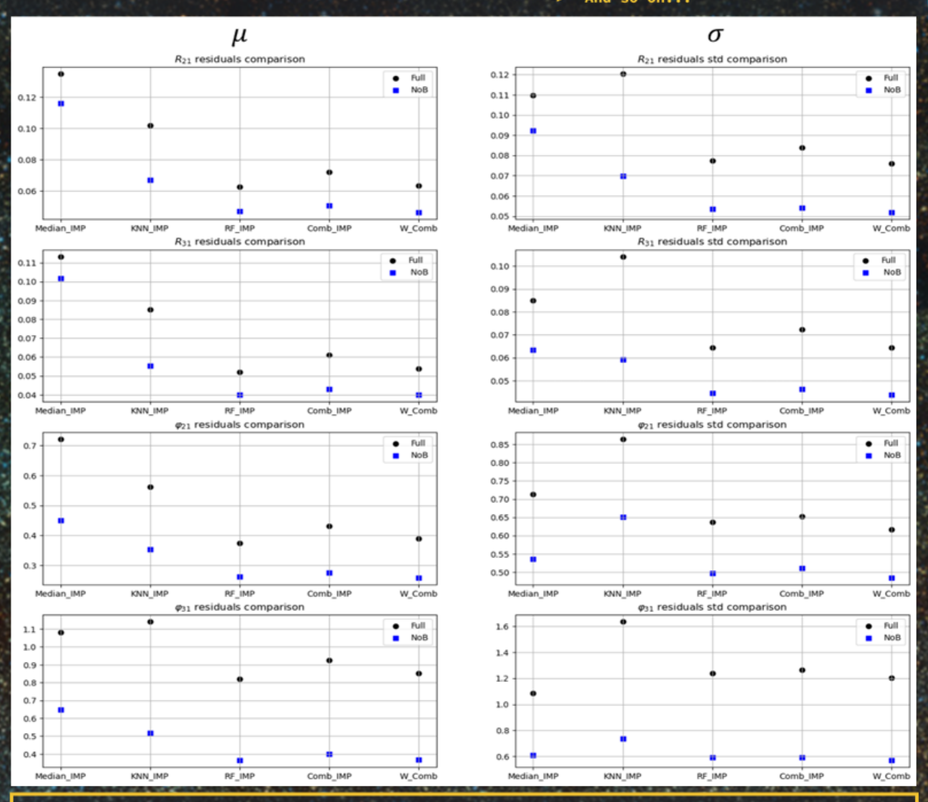
This is a preliminary study, in which the first given dataset presented the following problems:

- Limited amount of data samples, with a highly contaminated training set
- Minimal parameter space
- Highly unbalanced classes
- Heavy presence of duplicates

MISSING DATA IMPUTATION

The *Iterative Imputer* uses other features to predict (using modelling) missing values in a feature

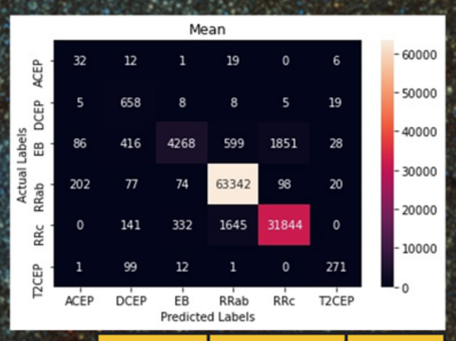




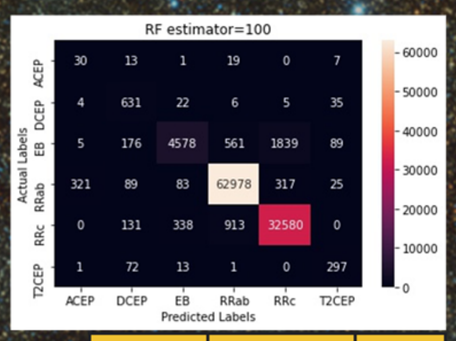
Above is the comparison, for the complete blind test sample, between each feature's mean and standard deviation residuals for five different imputation models. Values for the complete dataset are marked in black ("Full"), while the blue ones ("NoB") represent the datasets without the *Eclipsing Binaries* ("EB"), just to show how these objects are the main contaminants.

From left to right on x axis: imputation with the median, imputation with the sk-learn *KNNImputer*, the *Iterative Imputer* with the *Random Forest* regressor and the combined values of *Random Forest* and *KNN*, firstly as a simple arithmetic mean ("Comb_IMP") and lastly as a weighted mean ("W_Comb"; RF weight = 0.75, KNN weight = 0.25).

The plots clearly indicate that the Random Forest model is the best overall.



	purity (%)	completeness (%)	support
ACEP	9.82	45.71	70
DCEP	46.90	93.60	703
EB	90.91	58.89	7248
RRab	96.54	99.26	63813
RRc	94.22	93.76	33962
T2CEP	78.78	70.57	384



	purity(%)	completeness (%)	support
ACEP	8.31	42.86	70
DCEP	56.74	89.76	703
EB	90.92	63.16	7248
RRab	97.67	98.69	63813
RRc	93.78	95.93	33962
T2CEP	65.56	77.34	384

Here the confusion matrices and statistics show the blind test sample's classification performance comparison between a *Simple* and *Iterative* imputation, respectively using the mean and a *Random Forest* regressor with 100 trees.
For the final classification, both the knowledge base and blind test sample were imputed.

Focusing on completeness, in particular for the RR Lyrae, it can be seen that with *Iterative Imputer*, for "RRab" degradation is minimal, while for "RRc" we obtained an improvement.

THE DATA

	Gaia_id	period	ampG	R21	R31	phi21	phi31	Class
0	5878239339655855104	7.6	0.2	0.2	NaN	4.9	NaN	DCEP
1	4063826557167149440	5.7	0.5	0.4	0.1	4.8	2.9	DCEP
2	4060985694070626816	17.3	0.8	NaN	NaN	NaN	NaN	T2CEP
3	5254404136047673728	2.8	0.2	0.1	NaN	4.5	NaN	DCEP
4	486834100624652800	4.0	0.4	0.4	0.2	4.5	2.8	DCEP

Every object has an ID and a Class (ACEP, DCEP, EB, RRab, RRc and T2CEP)

The parameter space consists of 6 features:

- "period": period in days of magnitude variation
- "ampG": G band peak-peak amplitude
- "R21", "R31", "phi21" and "phi31": respectively amplitude ratios and phase differences obtained from the Fourier series of the light curves.

DATA		Nan FRACTION (%)					
FILE	TOTAL	period	ampG	R21	R31	phi21	phi31
knowledge base	18000	0	0	18.1	57.8	18.1	57.8
blind test	115000	0	0	9.6	57.2	9.6	57.2

DATA		Nan FRACTION (%)					
FILE	TOTAL	period	ampG	R21	R31	phi21	phi31
knowledge base	15354	0	0	17.6	57.1	17.6	57.1
blind test	106180	0	0	9	50.8	9	50.8

Upper left Table: NaN fraction for each feature in the original datasets, including duplicates.
Upper right Table: NaN fraction for each feature in the original datasets after dropping duplicates.

DATA	TOTAL	CLASSES					
		RR Lyrae		Cepheids			Binaries
		RRab	RRc	ACEP	DCEP	T2CEP	EB
Valid sources in knowledge base	6479	2564	484	156	1485	516	1274
Valid sources in blind test	52233	42322	7576	51	409	179	1696

DATA	TOTAL	CLASSES					
		RR Lyrae		Cepheids			Binaries
		RRab	RRc	ACEP	DCEP	T2CEP	EB
Sources in knowledge base	15354	3880	2071	209	2568	1230	5396
Sources in blind test	106180	63813	33962	70	703	384	7248

Classes distribution, following the duplicates removal, for the benchmark datasets (without NaNs) on the left and the complete datasets (with NaNs) on the right.
The knowledge base constitutes the training set for the classification, while the blind test sample is the target to classify.
From the tables above one can immediately notice how heavily unbalanced the classes are and how more than a half of both datasets contain missing data: the former problem is marginally improved by including the NaNs, while the latter was dealt with by adopting an imputation strategy.

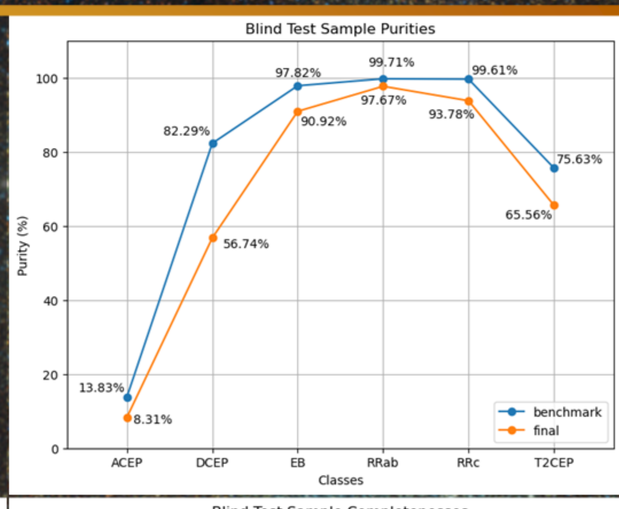
RESULTS

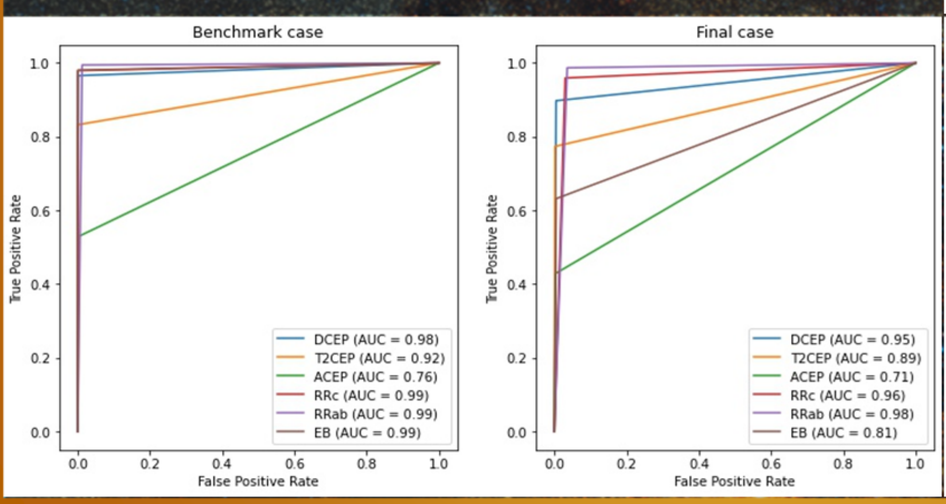
On the right the comparisons between each class purities and completenesses are shown, between the benchmark (no imputation) and final (with imputation) cases, for the blind test sample.

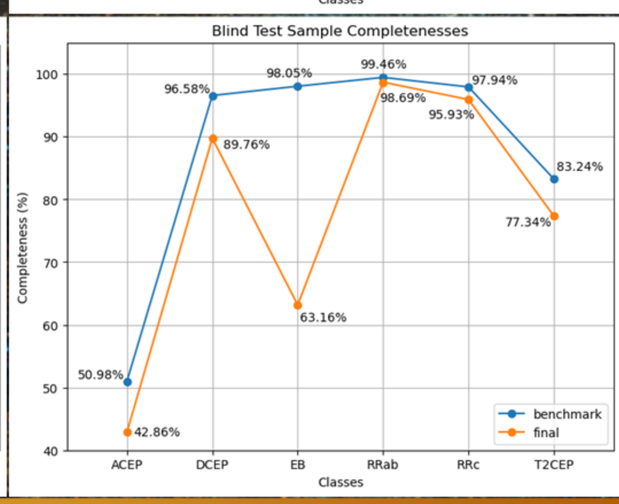
Apart from the Eclipsing Binaries ("EB") losing in completeness, all the other classes show minimum degradation and overall good performances similar to the benchmark best case scenario.

In the ROC curves below, the degradation for each class can be seen in more detail. In particular, evaluating the difference between the final and benchmark AUC values, one obtains:

$$\Delta RRab = -0.01$$
$$\Delta RRc = -0.03$$







CONCLUSIONS

- The benchmark case showed that *Random Forest* obtains good classification performances
- Comparison with benchmark case showed that the *Iterative Imputer* method gives satisfying results with minimum degradation, as seen in the ROC curves
- Imputation of missing data with the *Iterative Imputer* is more robust than using the mean

References:

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.

Scikit-Learn: Iterative Imputer (https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html)

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Image credit: ESO/VVV Survey/D. Minniti

INAF ISTITUTO NAZIONALE DI ASTROFISICA