Estimation of Stellar Parameters for S-PLUS DR2 stars based on Random Forests

Marcos Vinicius Emanuel Cordeiro da Silva, Lethycia Maria de Carvalho, Marcelo Borges Fernandes, Carlos Andres Galarza Arévalo, Simone Daflon, Raimundo Lopes de Oliveira Filho, Felipe de Almeida Fernandes, Cláudia Mendes Oliveira

16th S-PLUS Collaboration Meeting

Maceió – 02/12/21

Introduction

Blue Stars in the Halo

Observatório Nacional

The main objective of the project is the study of **Blue Stars in the Galactic Halo** using mainly the data provided by **S-PLUS**:

- Formed by **Hot Subdwarfs**, Horizontal Branch (HB) Stars, **post-AGBs**, Cataclysmic Variables (CVs), **Symbiotic Stars**, Planetary Nebulae and **Blue Stragglers**;

Blue Stars in the Halo

Observatório Nacional

The main objective of the project is the study of **Blue Stars in the Galactic Halo** using mainly the data provided by **S-PLUS**:

- Formed by **Hot Subdwarfs**, Horizontal Branch (HB) Stars, **post-AGBs**, Cataclysmic Variables (CVs), **Symbiotic Stars**, Planetary Nebulae and **Blue Stragglers**;
- Due to the **small number** of identified blue stars in the galactic halo, not a lot is known about their **formation mechanisms** and their **physical parameters;**

Blue Stars in the Halo

Observatório Nacional

The main objective of the project is the study of **Blue Stars in the Galactic Halo** using mainly the data provided by **S-PLUS**:

- Formed by **Hot Subdwarfs**, Horizontal Branch (HB) Stars, **post-AGBs**, Cataclysmic Variables (CVs), **Symbiotic Stars**, Planetary Nebulae and **Blue Stragglers**;

- Due to the **small number** of identified blue stars in the galactic halo, not a lot is known about their **formation mechanisms** and their **physical parameters;**

- Feeding the high-quality data from S-PLUS to Machine Learning (ML) algorithms, it's possible to train models capable of **identifying** (Classificators) these stars and also **estimating** (Regressors) values for their **parameters**, with the latter being the focus of this initial work.

STEllar Parameter Estimator

With that in mind, the focus of this work was the development of ML models capable of receiving **stellar magnitudes** (and colors) **as input** and returning a certain **stellar atmospheric parameter** (Teff, logg or [Fe/H]) **as output**:

Input

	mag 1	mag 2	mag 3	mag 4	mag 5
star 1	20.48	19.84	19.16	18.98	18.56
star 2	20.34	19.95	19.45	18.93	18.81
...	...	...	...	...	...

STEPE

Output

	Teff
star 1	5000
star 2	4800
...	...

Datasets

Input Features

The set of input features chosen for our models were 12 magnitudes given by **S-PLUS DR2**, 3 magnitudes given by **GAIA EDR3** and 4 magnitudes given by **WISE All-Sky Release**:

Filtro	Survey	CW (nm)
u	S-PLUS	348.5
J0378	S-PLUS	378.5
J0395	S-PLUS	395.0
J0410	S-PLUS	410.0
J0430	S-PLUS	430.0
g	S-PLUS	480.3
BP	GAIA	505.0

Filtro	Survey	CW (nm)
J0515	S-PLUS	515.0
G	GAIA	623.0
r	S-PLUS	625.4
J0660	S-PLUS	660.0
i	S-PLUS	766.8
RP	GAIA	773.0
J0861	S-PLUS	861.0

Filtro	Survey	CW (nm)
z	S-PLUS	911.4
W1	WISE	3352.6
W2	WISE	4602.8
W3	WISE	11560.8
W4	WISE	22088.3

As well as the **19 magnitudes** above, we also calculated all the **171** possible **colors** (difference between two magnitudes), resulting in **190 total input features**.

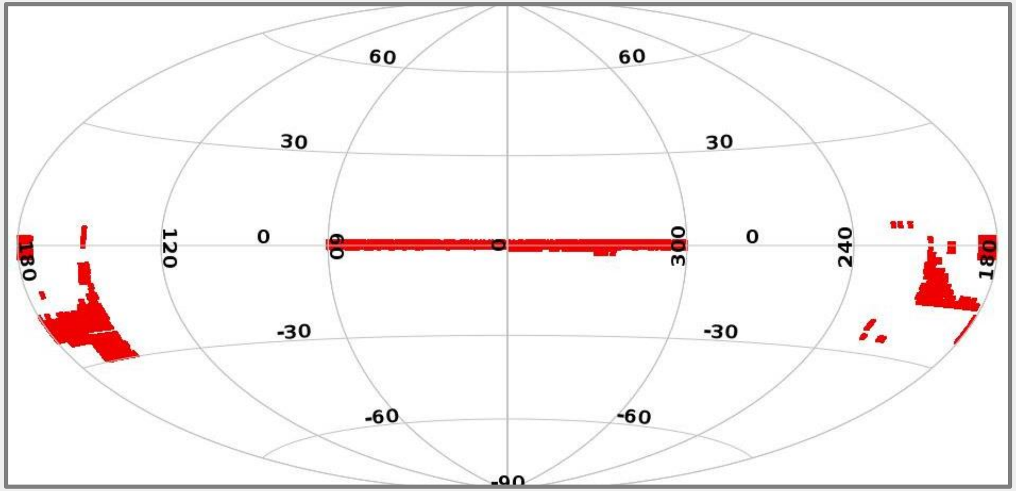
Input Features

Observatório Nacional

To ensure that we are working with **high-quality data**, the stars are **filtered** according to the following conditions:

- **prob_star** > 0.9
- **max(mag_err)** < 0.2
- **flag** = 0

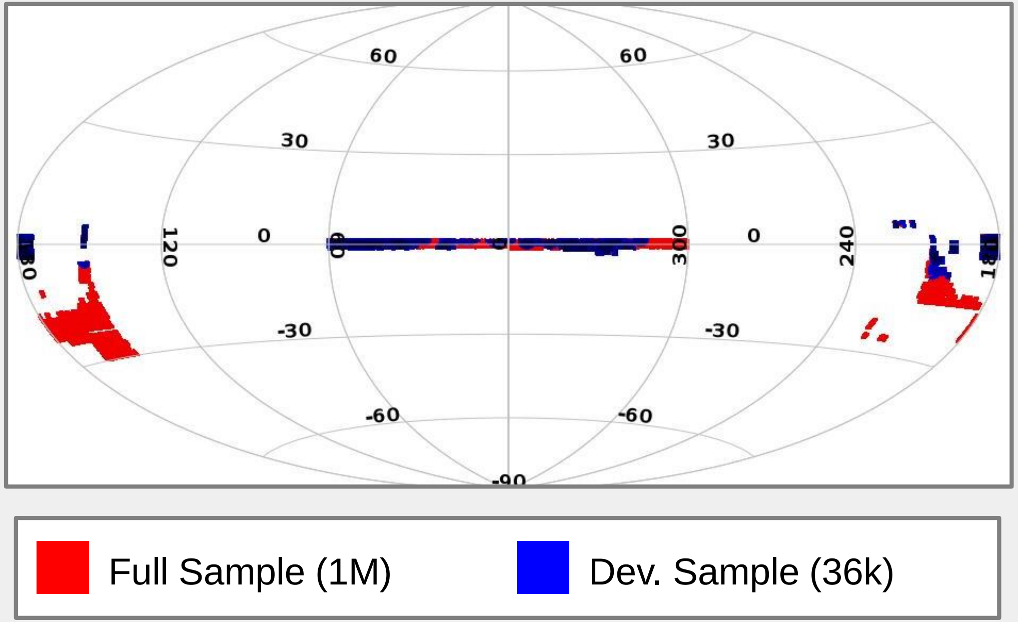
Our final dataset of stars observed by S-PLUS, GAIA and WISE has around **1 million stars** of all types, **not just Blue Stars**.



Full Sample (1M)

Development Sample

To develop our models, we need a subsample of stars with parameters already measured. To create that, we **cross-match** our S-PLUS/GAIA/WISE sample with the **LAMOST** DR6 survey data.



Again, to ensure the quality of the data, we apply some filters:

- **teff_err <** 300K,
- **logg_err** < 0.4,
- **feh_err** < 0.4.

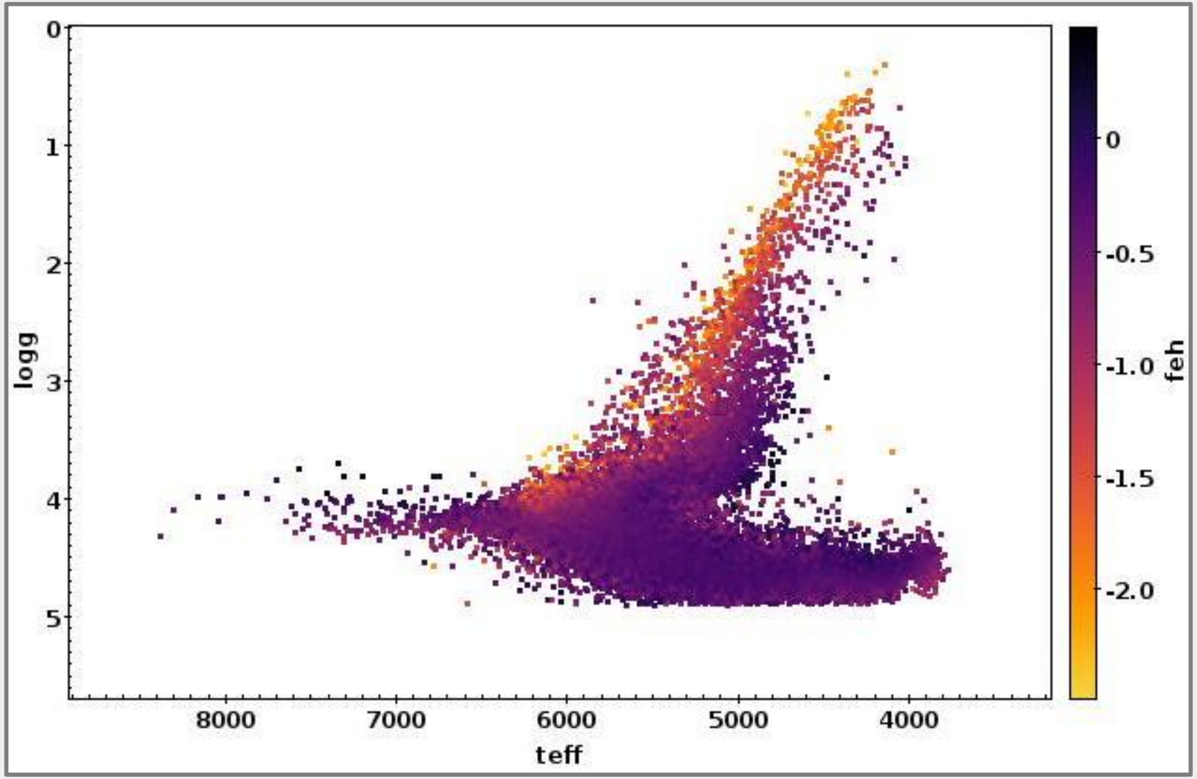
From the 36k resulting stars, we create the training (27k) and testing (9k) samples

Development Sample

As we can see, the development sample has stars with parameters in the following ranges:

- **teff**: [3800, 8000] [10k, ...]
- **logg**: [0, 5] [0, 6.5]
- **feh**: [-2.5, 0.5] [-2.5, 0.5]

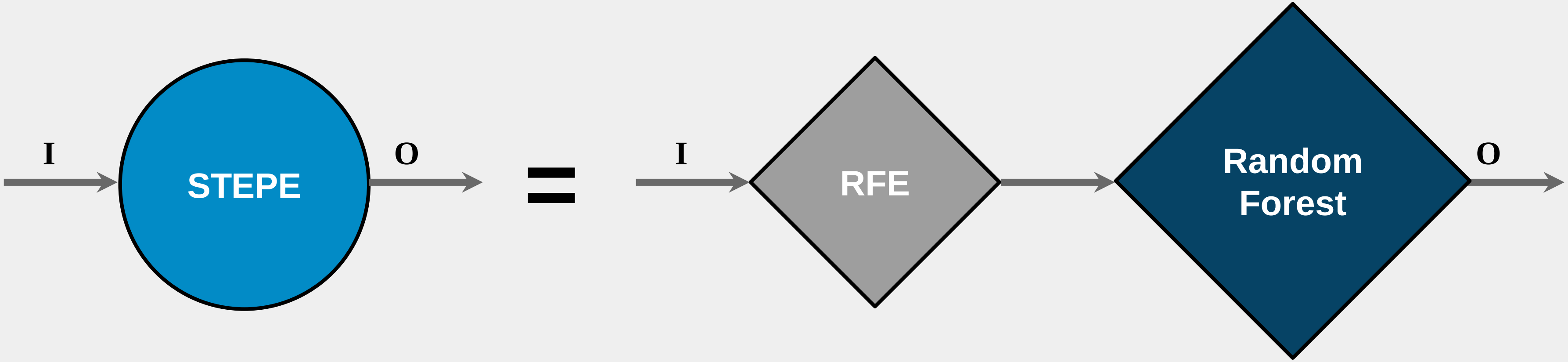
These intervals will define the **effective ranges** of our models.



Model Development

Model Structure

All of our models follow the same general structure, consisting of two steps:



1. **Recursive Feature Eliminator (RFE)**: Receives all the 190 input features and eliminates the worst ones, passing only the *n_features* best ones to the next step;
2. **Random Forest (RF)**: Receives the *n_features* best features from the last step and uses them to estimate the stellar parameter.

Hyperparameters

Inside each one our models, there’s a group of **hyperparameters (HPs)** that need to be chosen before any training can be performed. Among them, the **most important** are:

- **n_features**: Number of features that the RFE passes to the Random Forest estimator;
- **n_trees:** Number of trees in the RF;
- **max_features:** Fraction of features that each decision tree inside the RF considers when doing its splits;
- **min_samples_leaf (msl)**: Minimum number of objects on each side of a split for it to be considered valid.

Although it is possible to train our models with the default values of these hyperparameters, there’s **no reason** to believe that this combination is the best one.

Hyperparameters

With that in mind, we will **tune** our three estimators (Teff, logg and feh) **separately**, and try to find the best HP combination for **each one**.

In our case, the values to be tested are shown in the table above, and the evaluation method chosen is a **4-fold, 2-repeat cross-validation**, after which the R2-Score of each model is compared.

Hyperparameter	Values tested
n_features	[15, 45, 60, **190**]
n_trees	[50, **100**]
max_features	[0.25, 0.5, 0.75, **1.0**]
min_samples_leaf (msl)	[**1**, 10]

64 combinations for each STEPE

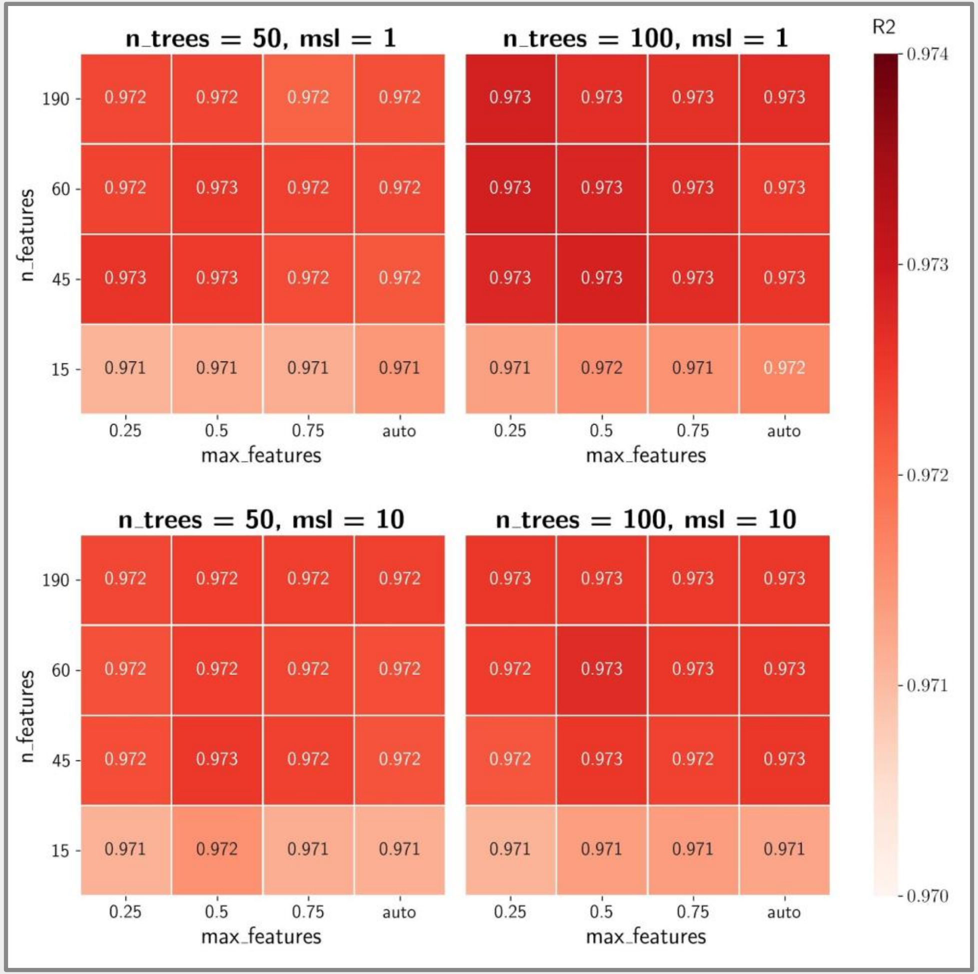
Hyperparameter Tuning

Results

STEPE - Teff

HP Combination	R2
(60, 0.25, 100, 1)	0.9729 ± 0.0012
(190, 0.25, 100, 1)	0.9729 ± 0.0015
(45, 0.5, 100, 1)	0.9729 ± 0.0018

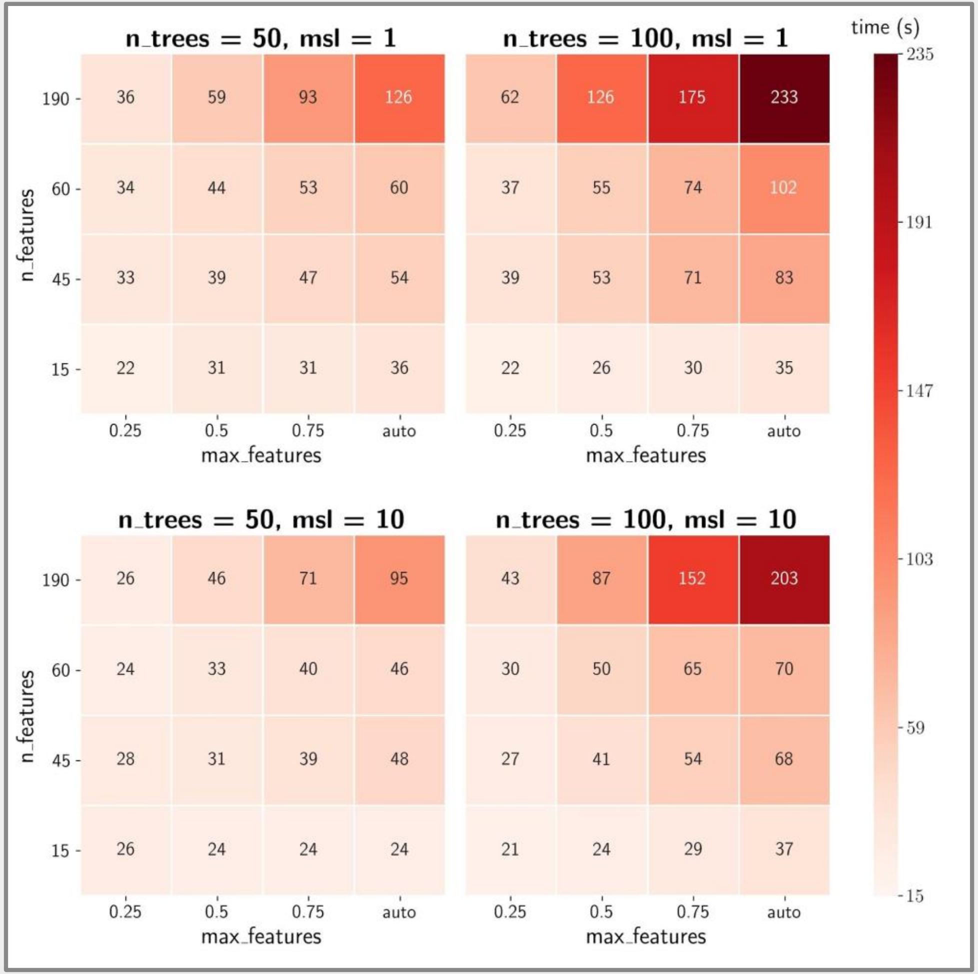
- Almost **no difference** in the score for values of *n_features* larger than 45;
- A R2-score of **0.9729** is equivalent to a **correlation of 98.63%** between the real value and the estimation.



STEPE - Teff

- In addition to not increasing the performance significantly, values of *n_features* larger than 45 also generate models with **much larger training times**;

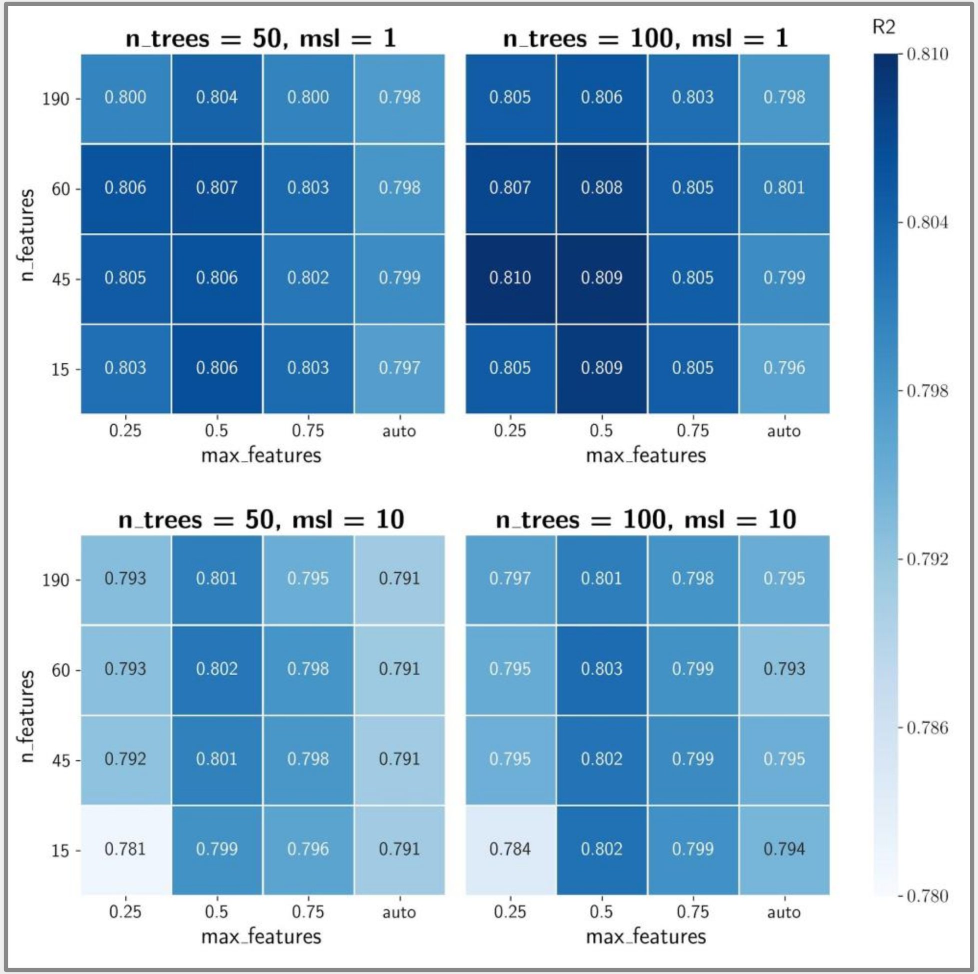
- Despite lowering the score of the models, a value of *msl* = 10 also **lowers the training time**.



STEPE - logg

HP Combination	R2
(45, 0.25, 100, 1)	0.8096 ± 0.0120
(45, 0.5, 100, 1)	0.8095 ± 0.0047
(15, 0.5, 100, 1)	0.8087 ± 0.0123

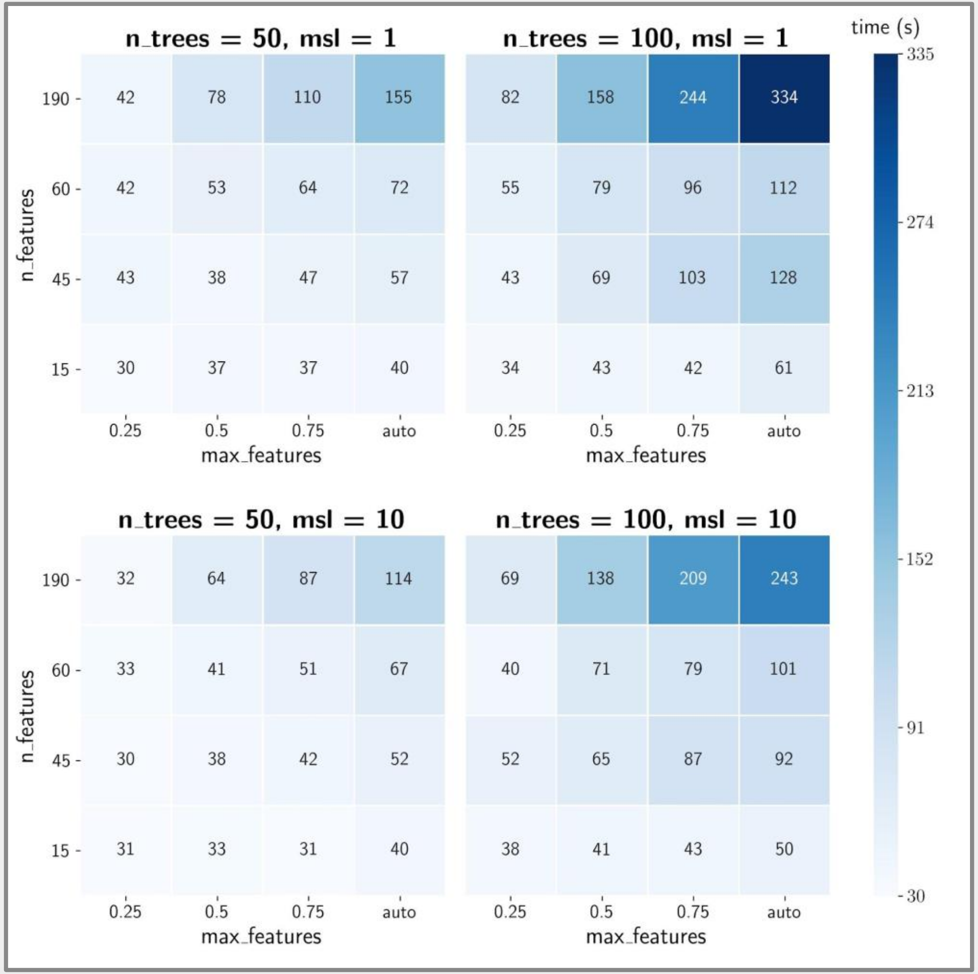
- The performance **peaks at** ***n_features* = 45** and becomes smaller for *n_features* = 60 and 190;
- A R2-score of **0.8096** is equivalent to a **correlation of 89.98%** between the real value and the estimation.



STEPE - logg

- In addition to decreasing the performance, values of *n_features* larger than 45 also generate models with **much larger training times**;

- Despite lowering the score of the models, a value of *msl* = 10 also **lowers the training time**.



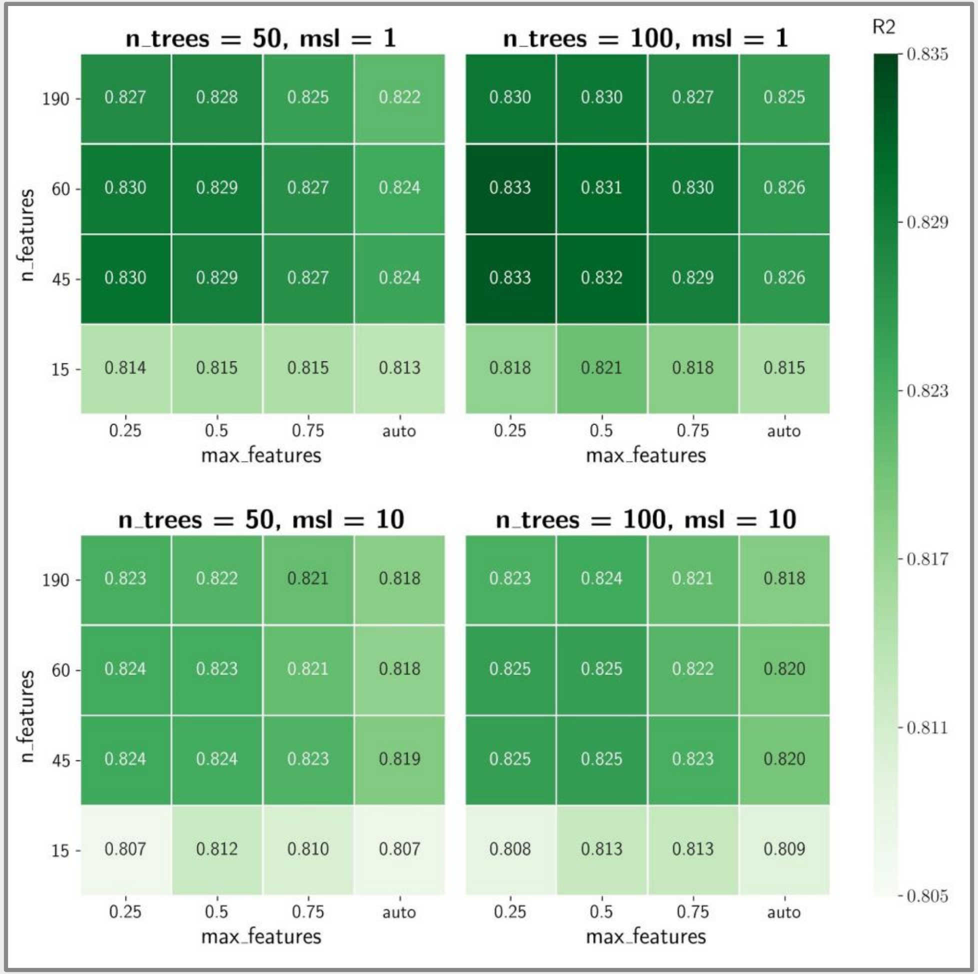
n_features	0.25	0.5	0.75	auto
n_trees = 50, msl = 1				
190	42	78	110	155
60	42	53	64	72
45	43	38	47	57
15	30	37	37	40
n_trees = 100, msl = 1				
190	82	158	244	334
60	55	79	96	112
45	43	69	103	128
15	34	43	42	61
n_trees = 50, msl = 10				
190	32	64	87	114
60	33	41	51	67
45	30	38	42	52
15	31	33	31	40
n_trees = 100, msl = 10				
190	69	138	209	243
60	40	71	79	101
45	52	65	87	92
15	38	41	43	50

STEPE - [Fe/H]

HP Combination	R2
(60, 0.25, 100, 1)	0.8331 ± 0.0034
(45, 0.25, 100, 1)	0.8330 ± 0.0037
(45, 0.5, 100, 1)	0.8318 ± 0.0041

- Almost **no difference** in the score for values of *n_features* larger than 45;

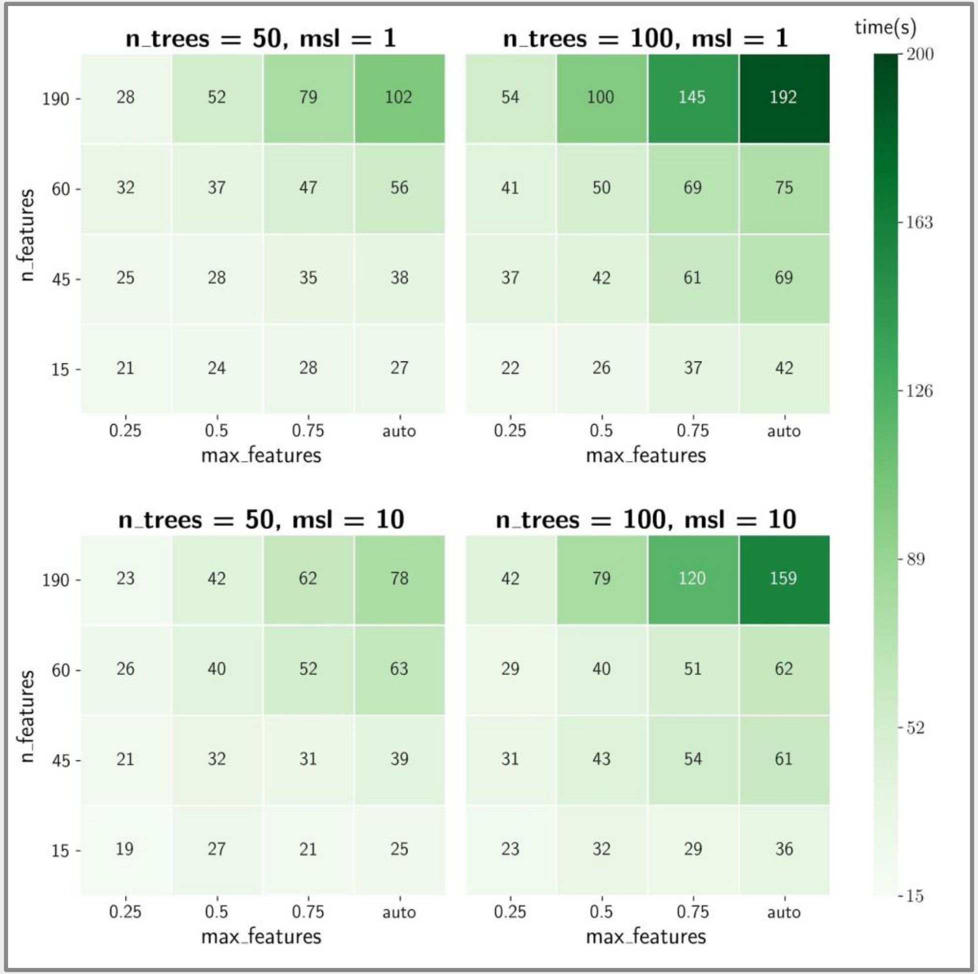
- A R2-score of **0.8331** is equivalent to a **correlation of 91.27%** between the real value and the estimation.



STEPE - [Fe/H]

- In addition to not increasing the performance significantly, values of *n_features* larger than 45 also generate models with **much larger training times**;

- Despite lowering the score of the models, a value of *msl* = 10 also **lowers the training time**.



Final Models

Hyperparameters

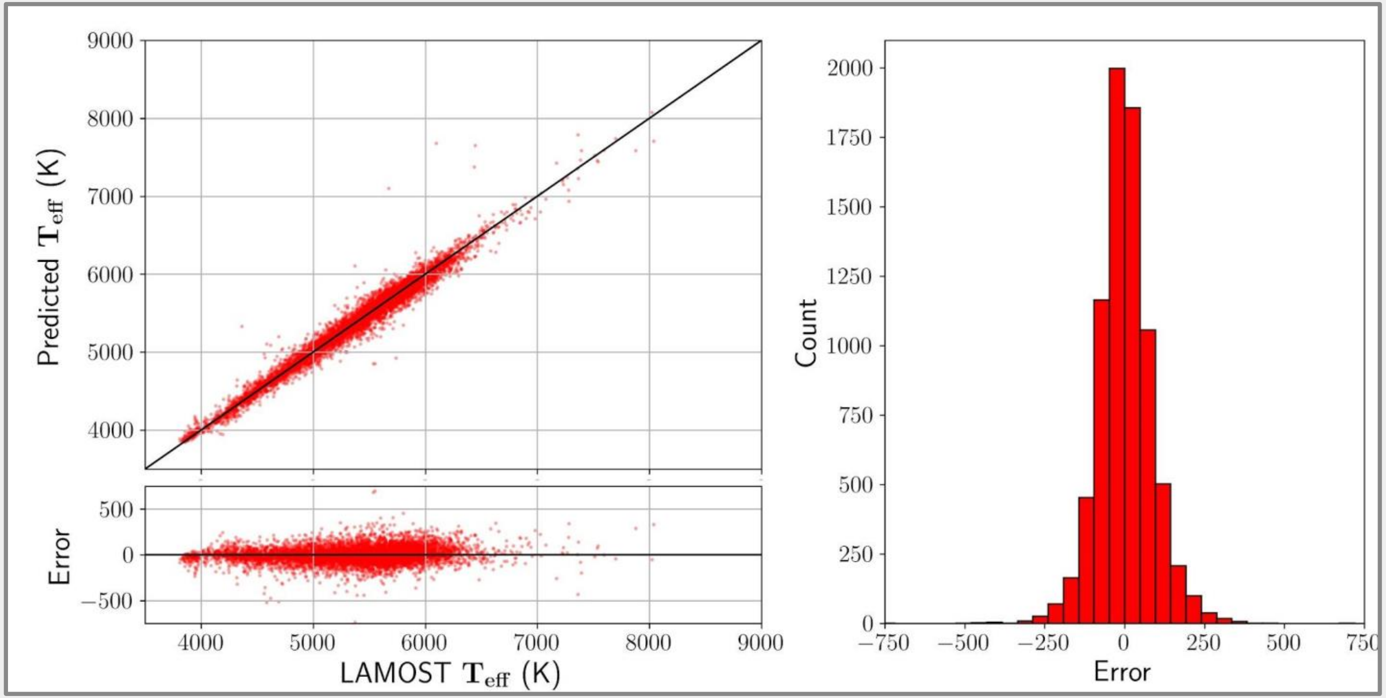
Observatório Nacional

The final combinations chosen for each STEPE were:

STEPE	n_features	max_features	n_trees	msl
Teff	60	0.25	100	1
logg	45	0.5	100	1
feh	60	0.25	100	1

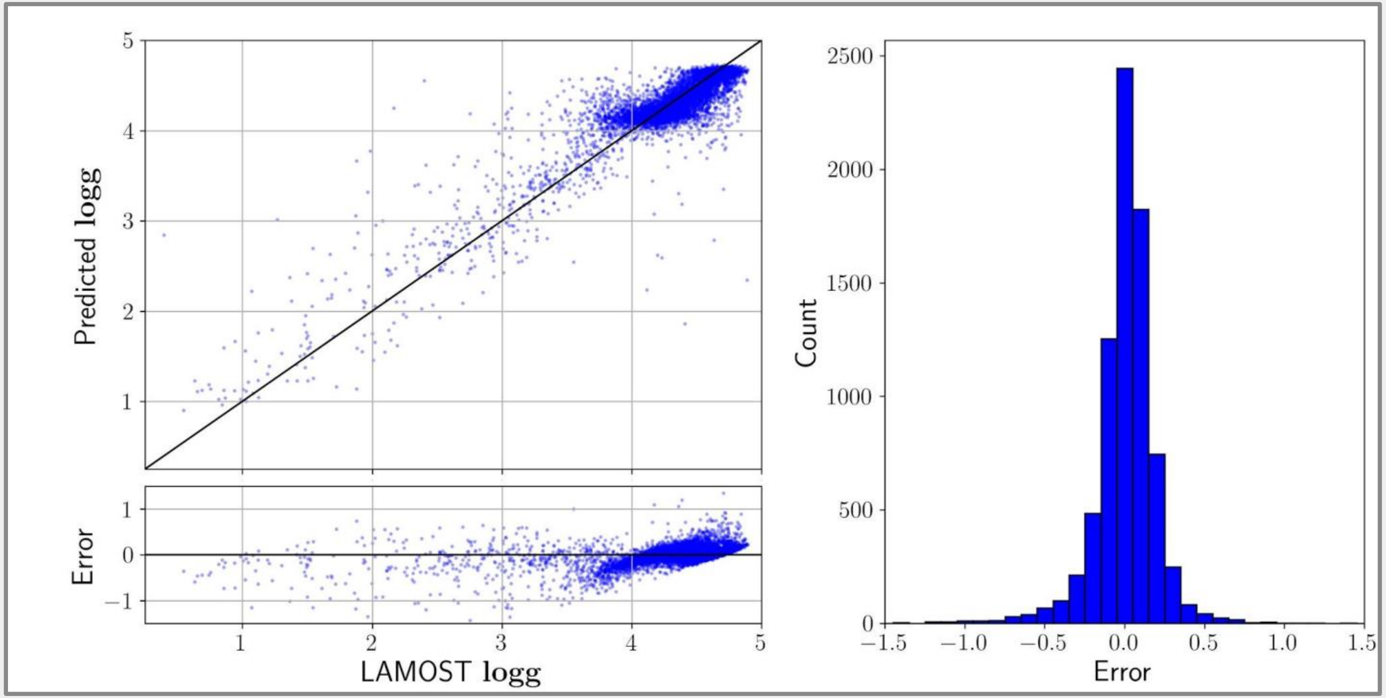
From here, the three final models were **trained on the whole training sample**, and their performance in new data was **evaluated on the initial test sample**, still not used up to this point.

STEPE - Teff



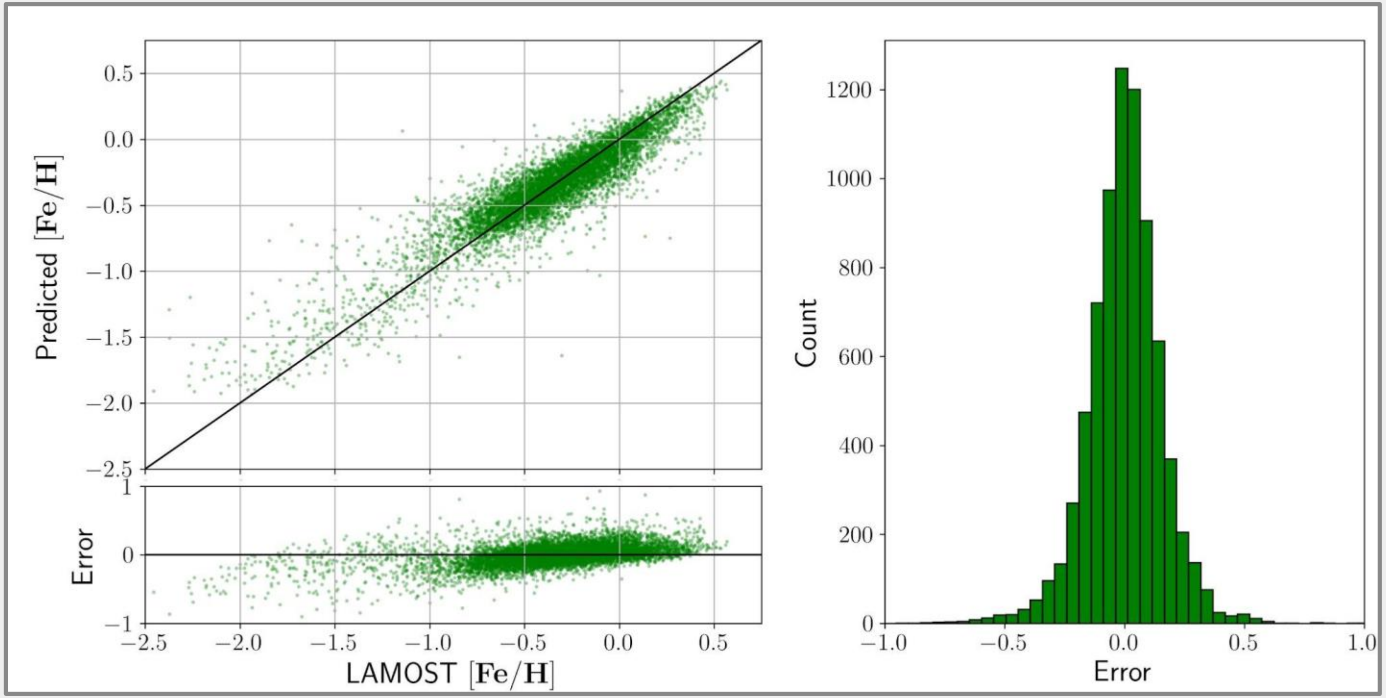
Metric	Value
MAE	63.992
RMSE	92.910
Max Error	1581.399
R2	0.973

STEPE - logg



Metric	Value
MAE	0.131
RMSE	0.208
Max Error	2.550
R2	0.819

STEPE - [Fe/H]



Metric	Value
MAE	0.115
RMSE	0.158
Max Error	1.334
R2	0.834

Conclusions

Observatório Nacional

Considering the results obtained from our hyperparameter tuning and final models, we can **conclude** that:

- A good portion of the 171 colors calculated **were not informative** for all three STEPEs, and their exclusion resulted in better models;

- The methodology used for the development of STEPEs showed **good results** and the final models were capable of giving **excellent estimations** for the stellar parameters in question when working with **general star data**;

Next Steps

Observatório Nacional

From here, there are some points that still need to be worked on in the future:

- **Expand** the effective range of the Teff STEPE (blue stars lay well outside the current range of 3800 - 8000 K);
- Test whether or not adding **more features** to the input data improves the performance of the model;
- Apply our methodology on samples **directed to blue stars;**
- **Compare** our results with the ones obtained by **other groups** working with Stellar Parameters and Machine Learning.

Thank You!

Development and Models: github.com/cordeirossauro/SPLUS-STEPEs

Contact: viniciuscordeiro@on.br

References

References

Observatório Nacional

Mendes de Oliveira, C., Ribeiro, T., Schoenell, W., Kanaan, A., Overzier, R. A., Molino, A., Sampedro, L., Coelho, P., Barbosa, C. E., Cortesi, A., Costa-Duarte, M. V., Herpich, F. R., Hernandez-Jimenez, J. A., Placco, V. M., Xavier, H. S., Abramo, L. R., Saito, R. K., Chies-Santos, A. L., Ederoclite, A., … Zaritsky, D. (2019). **The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies, and redshifts with 12 optical filters**. Monthly Notices of the Royal Astronomical Society, 489(1), 241–267.

Breiman, L. **Random Forests**. *Machine Learning*, 45, 5–32 (2001).

Refaeilzadeh P., Tang L., Liu H. (2009) **Cross-Validation**. In: LIU L., ÖZSU M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research, 12, 2825–2830.*

References

TensorFlow Developers. (2021). **TensorFlow (v2.6.2)**. *Zenodo*.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. **Array programming with NumPy**. *Nature*, 585, 357–362 (2020).

The pandas development team. (2020). **pandas-dev/pandas: Pandas 1.0.3 (v1.0.3)**. *Zenodo*.

Michael L. Waskom (2021). **seaborn: statistical data visualization**. *Journal of Open Source Software*, 6(60), 3021.

Hunter, J. (2007). **Matplotlib: A 2D graphics environment**. *Computing in Science & Engineering*, 9(3), 90–95.

Taylor, M. (2005). **TOPCAT & STIL: Starlink Table/VOTable Processing Software**. In *Astronomical Data Analysis Software and Systems XIV* (pp. 29).

References

Rijn, J., & Hutter, F. (2018). **Hyperparameter Importance Across Datasets**. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2367–2376).