Estimating spectroscopic ages of red-giant stars using machine learning

Author: Pol Gispert Latorre
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Friedrich Anders

Abstract: Over the last few years, many studies have found an empirical relation between the abundance of a star and its age, rather well known as chemical tagging. Here we estimate spectroscopic stellar ages for 197.000 stars observed by the APOGEE survey. To this end, we use the supervised machine learning technique XGBoost, trained on a set of 3314 stars with asteroseismic ages observed by both APOGEE and Kepler (Miglio et al. 2021). Eventually, to verify the obtained age estimates, we investigated the chemical, kinematic and positional relationship of the stars in respect to their age.

I. INTRODUCTION

Frequently, isochrone matching is used to determine stellar ages of the main sequence turn off and sub-giant branch. Another well-tested (but also model-dependent) method to estimate ages for field stars is asteroseismology. The main problem of these methods remains on the fact that they are not feasible nor accurate for an enormous sample of stars, just usable for a limited group of them. Therefore, for large-scale spectroscopic surveys like GALAH, other methods are necessary in order to know their age.

Here is where a recent study [1] comes into play, as it confirms an obvious relation between the age and abundances in field stars from the GALAH survey. This work hint at the possibility of Weak Chemical Tagging: the abundances of a star can be enough to determine its birth time, and possibly also its approximate birth position. In this way, we can study the kinematic and spatial structure of the galaxy (Galactic Archaeology), as it will be shown in this paper.

In addition to that, some authors [2] [3] have also considered the possibility of Strong Chemical Tagging: the idea to link the abundance pattern of a star directly with its birth cluster or association.

A recent study [4] shows that this is not realistic. This research explains that it is not clear if each cluster just has one chemical signature, if this one can be different from the neighbour clusters (overlapping chemical signatures) or even if these signatures evolve over time. In this study it has been found that more of the 70% of groups of stars marked by the chemical signature actually belonged to groups created statistically from stars pertaining to different real clusters. It is therefore unlikely to recover most birth clusters. Hence, dismissing the Strong Chemical Tagging, the Weak Chemical Tagging is used in our research.

Another study [5] found that approximately 30 elements (from the totality of those he resolved to study) show significant trends with age, a fact that suggests/ shows the path to follow. Thus, if abundances can be measured accurately, it is potentially possible to estimate the stars' age.

A similar procedure to the one used in [1] will be carried out, in which if the same algorithm is used, it studies different types of stars. This research will use the data test of Kepler field and will extrapolate the age-chemistry relations to the wider galaxy. Moreover, while [1] has utilised main-sequence turn-off stars, ours will be red-giant stars.

So, the objectives of this work is reproducing the cinematic and spatial measures with the predicted chemical ages. These chemical ages will be predicted with an algorithm of Extreme Gradient Boosting using as training the data from Kepler field and extrapolating the model to new zones of the galaxy.

II. EXPERIMENTAL

A. Machine Learning basics

A brief introduction about the basic functioning of a machine learning model will be shown in the following lines. A machine learning model is a mathematical representation of a system which learns about the introduced data. This models are used to predict outcomes based on an introduced input (a feature). The features are the variables of the data that the model will utilise to train itself. So, it will be necessary to introduce input features in order to obtain the corresponding labels. Therefore, when the model is trained, it can be used to make predictions on new data. We can see a visual schema of this explanation in the Figure 1.

B. XGB Algorithm

The Extreme Gradient Boosting will be used in this paper. This algorithm is the culmination of other machine learning algorithm development. The basis of this model is a decision tree, which is a graphical representation of possible solutions to a decision based on certain conditions. Above this basic machine learning model was established the random forest, which is a bagging-based

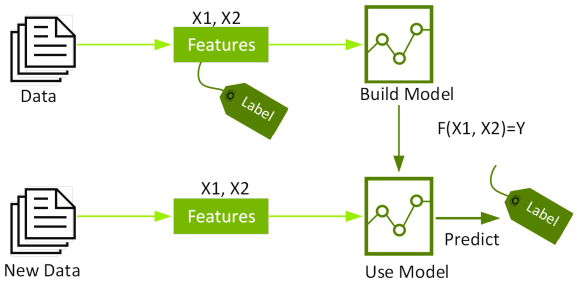

FIG. 1: Schema of a simple machine learning system. The symbol marked as "Data" corresponds to our trained set of 3.314 stars and the one marked as "New Data" corresponds to the 197.000 stars from APOGEE survey.

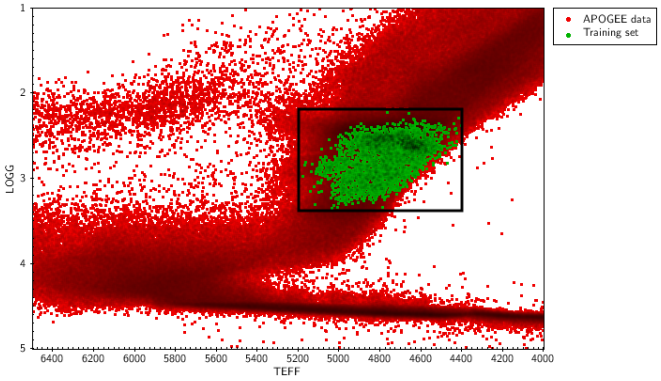

FIG. 2: Kiel diagram (log g vs. Teff) for the APOGEE DR17 catalogue (red density). In green our overplot the training sample (taken from the APOGEE-Kepler catalogue of Miglio et al. 2021). The box highlights the stellar parameter range for which our method provides stellar age estimates.

algorithm where we select randomly the features to construct a decision tree. Subsequently, due to the previous models, the Boosting and Gradient Boosting methods were created, which minimise the errors and increase the performance of the models. This last model is different due to the fact that it uses a gradient descent algorithm to minimise the errors of the sequential models.

Finally, taking over from the previously described algorithms, XGB differs from gradient boosting in that it performs parallel processing, tree-pruning, handling missing values and regularises the data to avoid over fitting [6]. Also, his training time is incredibly minor compared to the gradient boosting and random forest, and whose prediction power is similar to gradient boosting.

We will make a brief explanation of how the algorithm works [6].

The overall optimisation objective for XGB is to search the minimisation of the regularised loss function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t) \qquad (1)$$

where $\mathcal{L}$ is the loss function, y_i is the true label of the i-th iteration, $\hat{y_i}$ is the predicted label and $\Omega(f_t)$ is the regularisation term. The model parameters are given by t. The loss function is similar to other machine learning functions as logistic loss or the least squares loss.

The regularisation term $\Omega(f_t)$ is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \qquad (2)$$

where γ is a parameter that controls the complexity of the model, T is the number of leaves in the tree and w_j is the weight of the i-th leaf. The regularisation term is used to prevent over fitting by adding a penalty term to the loss function.

C. Data selection

Once the prediction algorithm has been explained, the data selected to train the model are limited at $4400<TEFF<5200$ and $2.2<LOGG<3.4$ of all APOGEE data. In figure we can see a plot of all the data of the survey and our selected training data.

The area chosen to be studied are 3314 stars preceding from the Kepler Field [7]. It is necessary to differentiate the red clumps from the rest of the stars, since due to the CNO cycle it will have a different metallicity, so that their chemical ages can be better predicted.
The chosen features to train the model are:

- effective temperature (Teff)
- logg
- chemical abundances

The chosen chemical abundances are: [C/Fe], [Cl/Fe], [N/Fe], [O/Fe], [Na/Fe], [Mg/Fe], [Al/Fe], [K/Fe], [Ca/Fe], [Ti/Fe], [V/Fe], [Mn/Fe], [Co/Fe], [Ni/Fe]. The other abundances are not used because we discarded them using the SHAP values method.

D. SHAP values

The SHAP (SHapley Additive exPlanations) values provide information to understand the output of a machine learning model [8]. In XGB's case, they can be used to understand how every feature has an impact on the predictions of the model.

The sum of SHAP values is equal to the difference between the expected output and the baseline output. In other words, a positive value means that the feature increases the output's model, while a negative value decreases it.

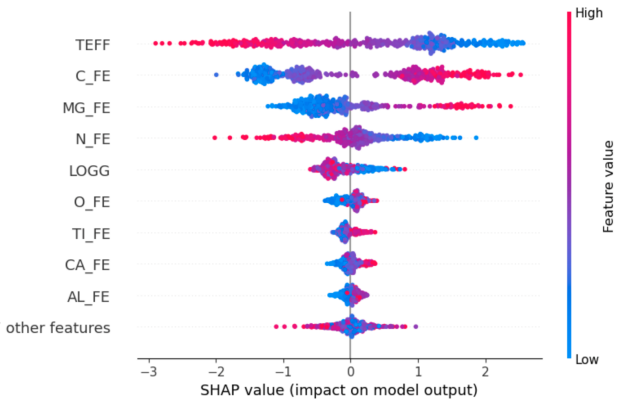

FIG. 3: Representation of the SHAP values of our trained model, where the X-axis represent to the SHAP value, the Y-axis corresponds to the different labels and the colour indicates the magnitude of the feature value.

In the present case, the SHAP values has been used not only to recognize the metallicites which have a major impact on the model, but also to discard those who don't contribute to the prediction (data selected in the section II.C).

The SHAP values with major impact in the model (Fig 3.) are the fffective temperature and the chemical abundances [C/Fe], [Mg/Fe] and [N/Fe].

To understand how it works we will make an example: if we look at the effective temperature, we can observe that an increase in its value (high feature value) implies a negative shap value, i.e. a negative impact on the model. This implies that the predicted age decreases. It makes sense since the higher the temperature, the lower the age of the stars and vice versa.

Thanks to this method we were able to discard the following chemical abundances because their shap value was too low: [Si/Fe], [P/Fe], [S/Fe], [Cr/Fe], [Fe/H], [Cu/Fe] and [Ce/Fe].

E. Grid Search

Grid Search [9] is a method for hyper-optimising parameters for a parameters for a machine learning model. It is based that, by defining ourselves a set of possible parameter values, the algorithm will train and evaluate models using all possible combinations of the parameter values, so it will find the best set of parameter values for the model. We use it because gives us the best performance of the model.

The optimal parameters obtained are:

- learning rate: 0.03
- max depth: 5
- min child weight: 4
- n estimators: 500

FIG. 4: Comparison of the predicted age (y-axis) vs. test age (x-axis). The diagonal line represents perfect prediction. The scatter plot illustrates the accuracy of the model, with denser clusters around the diagonal line indicating better performance. The colour of the points represents the evolutionary state of the star

- ntheard: 4
- objective: "reg:linear"
- silent: 1
- subsample: 0.7

These parameters have been used in the final modelling to optimise the results of the regression.

III. RESULTS AND DISCUSSION

A. Model training results

We will briefly show the performance of the model before showing the results of its application to the large table to be studied.

If we compare the chemical ages obtained from the seismic ages (Fig. 4) it is observable a good correlation of $R^2 = 0.91$

Our XGBoost regression is capable of estimating meaningful ages for both red-clump stars (blue points in Fig. 4) and first-ascent red-giant branch stars (red points in Fig. 4). We can appreciate that the performance for the red-clump stars is significantly better (lower dispersion around the identity line), but the model still works well on average for RGB stars. The statistical age error of our method is around 25%, almost independent of stellar age.

B. Chemical, spatial and kinematic relations with age

Our model has been applied to 197.000 APOGEE DR17 [10] stars located within the Teff - log g box highlighted in Fig. 1. To be sure that the results are correct, in the following subsections we show a number of validation plots that reproduce well the expected chemical, positional and kinematic trends with age.

1. *Age-Metallicity relations*

At the beginning of the Milky Way's formation, the Milky Way consisted of a halo of gas and the first stars began to emerge. These first stars, when they died, resulted in type II supernovae, in which oxygen and magnesium were released [11]. As the Milky Way formed into a thin disc, the number of type II supernovae decreased. These stars are now located in the thick disc, due to the halo in which the galaxy formed. So, the thick disk will have a high concentration of elements such as O and Mg formed in type II supernovae and its stars will be very old.

In contrast, as the thin disk formed, type Ia supernovae, in which Fe is released, increased. The thin disk formed later than the thick disk, so its stars must be younger. Therefore, it is to be expected that there are younger stars in the thin disk and older stars in the thick disk. And the ratio of magnesium to iron is lower in the thin disk and higher in the thick disk.

We can see this information reflected in the image above in Fig. 5 [12]. We have plotted our results in the image below in Fig5., with a similar axis ranges to the image above. We can see that we have obtained very good results, with a zone with older ages and another area with younger ones, corresponding to the tick and thin zones respectively. Also, it is accomplished the same relation with the [Mg/Fe] and [Fe/H] chemical abundances. The white contours in the lower panel of Fig. 5 represent the density iso-contours, so it is easier to see the two different zones.

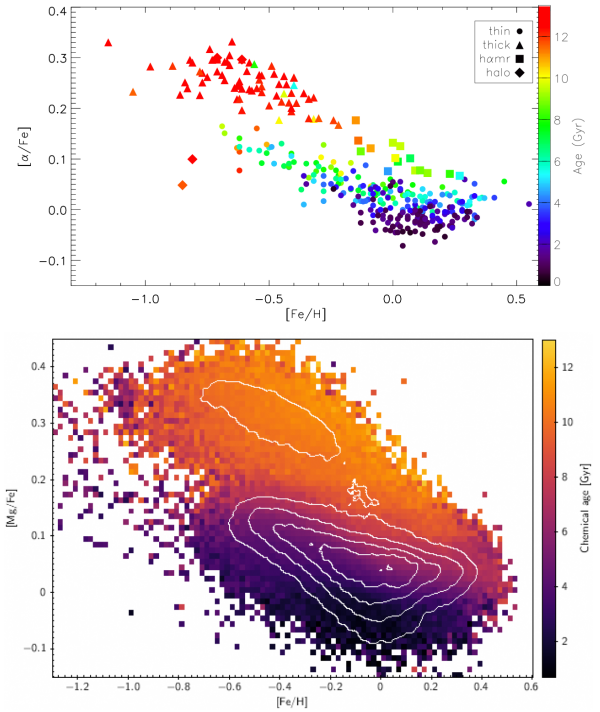

FIG. 5: [alpha/Fe] vs. [Fe/H] diagram, colour-coded by stellar age. Top panel: High-resolution solar-vicinity sample [12]. Bottom panel: Our results for the full APOGEE red-giant sample with spectroscopic ages.

2. *Age-Spatial relations*

In this section we have investigated how the age of stars changes as a function of the height above the Galactic plane (Z_{Gal}) and distance from the Galactic centre (R_{Gal}). We can see a plot in Figure 6.

We can see that for a radius smaller than 8-10kpc, it is true that for the thin disk the stars are younger and for the thick disk the stars are older. We can also see that the thickness of the thin disk in the plot (0.3 kpc approx.) agrees with the data we currently have on the Galaxy.

For a radius larger than 8-10kpc, we can observe that most of the stars are young. We know that the thick disk of the galaxy has a strong radial age gradient [13]. In which the median age for red clump stars ranges from 9 Gyr in the inner disk to 5 Gyr in the outer disk. This strong radial age gradient can be attributed to the formation process of the galaxy: the inner regions of the galaxy formed earlier faster than the outer regions. As a result, there is a higher concentration of older stars in the inner parts of the galaxy and higher concentration of younger stars in the outer parts.

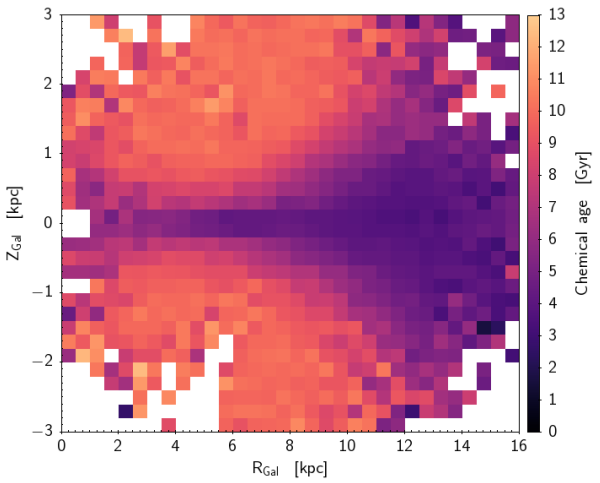

FIG. 6: Z_{Gal} vs. R_{Gal} diagram, colour-coded by stellar age. Stars from the full APOGEE red-giant sample.

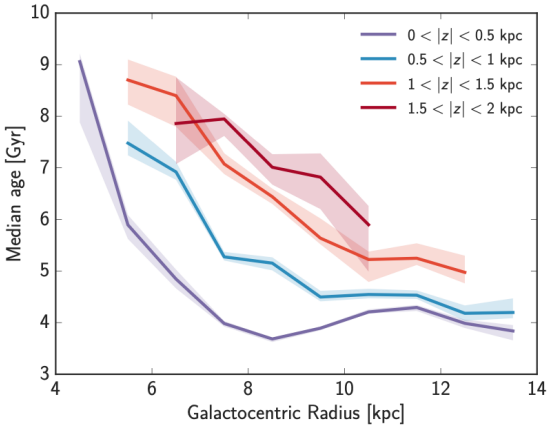

FIG. 7: Median age vs. galactocentric radius, colour-coded by height with respect to the galactic axis [13]. Stars from the outer parts of the Milky Way.

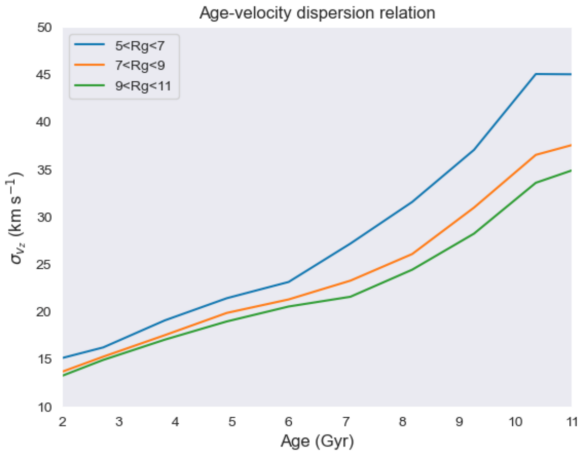

FIG. 8: Age-velocity dispersion relation within the ranges 5kpc to 11kpc, in various plots. The slope of the line represents the rate of change in the velocity dispersion with respect to chemical age.

Figure 7, obtained by [13] from earlier APOGEE data, illustrates the radial age gradients in the Galactic disc at different heights above the Galactic plane, which our Fig. 6 reproduces well at least qualitatively.

3. Age-Kinematics relations

Finally we compare the velocity dispersion of the stars with their chemical age. The data shown in Fig. 8 are for a galactocentric radius between 5kpc and 11 kpc, in various plots. The age-kinematics relations are accurately replicated when compared to [1]. But if we propose that the decrease in gravitational potential as we move away from the center of the galaxy leads to an increase in scattering velocity, this hypothesis is disproven. Consequently, we need further investigation in the age-kinematics relations.

IV. CONCLUSIONS

In summary, our XGboost algorithm has been able to accurately estimate the spectroscopic ages of the red-giant stars observed by the APOGEE survey. The results show a correct reproduction of the expected chemical, positional, and kinematic trends with age. This suggests that the Weak Chemical Tagging method is a valuable tool in galactic archaeological research, as it allows a large amount of information to be obtained from the chemical abundances.

In addition, it could be mentioned that the use of machine learning algorithms such as XGboost has allowed for greater efficiency and accuracy in estimating the spectroscopic ages of stars, which is especially important in large-scale studies such as GALAH. One could also point out the importance of continuing research in this field, as knowledge of the ages and detailed chemical abundances of stars can provide valuable information about the evolution of the Galaxy and its formation.

Acknowledgments

I want to express my gratitude to my advisor, Dr. Friedrich Anders, for his guidance and support throughout the project. I would also like to thank my family and friends for their support and encouragement.

[1] M. R. Hayden et al. 2022, MNRAS, 517, 4, 5325-5339
[2] Bland-Hawthorn 2002, Ann.Rev.Astron.Astrophys, 40, 487-537
[3] Price-Jones et al. 2020, MNRAS, 496, 4, 5101–5115
[4] L. Casamiquela et al. 2021, arXiv:2108.13431 [astro-ph.GA]
[5] Sharma et al. 2022, The GALAH Team, MNRAS, 517, 4, 5325–5339
[6] Tianqi Chen and Carlos Guestrin 2016, arXiv:1603.02754 [cs.LG]
[7] Miglio et al. 2021, arXiv:2004.14806 [astro-ph.GA]
[8] Scott M.Lundberg and Su-In Lee 2017, arXiv:1705.07874 [cs.AI]
[9] scikit-learn: Machine Learning in Pyhton, *https://scikit-learn.org/stable/modules/grid_search.html*
[10] Abdurro'uf et al. 2021, arXiv:2112.02026 [astro-ph.GA]
[11] Chiappini et al. 1997, Astrophys.J., 477, 765
[12] Delgado-Mena et al. 2019, A&A 624, A78
[13] M. Martig et al. 2018, arXiv:1609.01168 [astro-ph.GA]