ARTIFICIAL INTELLIGENCE FOR PREDICTIVE MODELING IN JEWELRY MANUFACTURING

Valentina Allodi, Francesco Gatto, Chiara Conti, Patrizio Sbornicchia, Valerio Doppio, Damiano Zito
Progold S.p.A.
Trissino (VI), Italy

ABSTRACT

In the field of jewelry manufacturing, the traditional characterization of gold and silver alloys relies heavily on extensive physical, chemical, mechanical, and technological testing. This research advances the state of the art by introducing an innovative approach using Artificial Intelligence (AI) and Big Data. Leveraging these technologies, we create predictive models for untested alloy compositions, forecasting critical properties such as melting range, color, hardness, among others. This technique significantly reduces the need for conventional testing, enhancing both accuracy and efficiency. The methodology encompasses data analysis, AI-driven predictions, and validation, focusing on numerous essential attributes in alloy design. Our work represents a pioneering contribution to the industry, opening new avenues for material understanding and streamlined production.

INTRODUCTION

Over the years, a significant amount of data has been accumulated through the analysis of our alloys. Recently, an exploration of the potential of Artificial Intelligence (AI) was initiated to see if AI could be used to predict the primary physical properties of future alloy compositions.

To undertake this endeavor, the methodology of machine learning was adopted[1,2]. Machine learning, a branch of artificial intelligence, specializes in training algorithms to recognize patterns in data sets and make predictions autonomously. Following a process of data refinement, efforts were directed towards optimizing predictive accuracy through a hybrid approach. This approach involved combining machine learning with experimental knowledge, for example, to consider only those elements in compositions that have a tangible impact on the characteristic under investigation.

A comprehensive evaluation was carried out to assess the robustness of the prediction results. This analysis included not only a review of relevant metrics, but also a comparative analysis of prediction errors versus actual experimental variances for each property under investigation.

WHAT IS MACHINE LEARNING?

Machine learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without being explicitly programmed.

A key-feature of machine learning is the machine ability to predict a behavior based on experience, namely input data, mimicking the way humans base decisions[3]. This self-learning process relay on the application of statistical modelling to detect patterns and improve performance based on data and empirical information. In other words, machine learning means to perform a set task using input data instead than input command.

This doesn't mean that the programmer is not involved in the learning process: feeding the data into the model, selecting the appropriate algorithm, and tuning its settings are actions carried out by the programmer that have a fundamental impact on the predictive results.

Machine learning is divided in three main categories[4]:

- Supervised Learning, where the model is trained on a set of annotated data, meaning data that includes both input features and corresponding desired output labels.
- Unsupervised Learning, where the model is trained on unlabeled data, and the goal is to discover hidden structures or patterns in the data.

- Reinforcement Learning, where the learning process involves training an agent to interact with an environment to achieve a goal. The agent learns to take actions that maximize some notion of cumulative reward. It learns through trial and error, receiving feedback in the form of rewards or penalties. Examples include training a robot to navigate a maze or teaching an AI to play a video game.

The learning processes utilized in this work involve data with input features linked to output values, placing them within the realm of supervised learning among the three subdivisions of machine learning.

TRAINING PROCESS

The first step of the training process is data cleaning. This might involve modifying or removing incomplete, irrelevant, incorrectly formatted, or duplicated data. This phase can often be the most time-consuming of the entire process. In the specific cases that were examined in this study, the data cleaning phase primarily involved checking that the data recorded in the past were extracted correctly from our database, and secondarily, identifying and correcting some errors present in the original database itself.
After data cleaning, there is typically a data pre-processing phase, which can include data normalization and standardization. Following this, the input data is divided into training and test sets. The training data set is used to develop and optimize the model, while the test data set is used to evaluate the model's performance in prediction. It is crucial to ensure that the division between training and test data does not inadvertently omit significant variance from the training data, as unexpected surprises may arise when applying the trained model to the test data.
After the splitting into training and test sets is completed, it is time to train the model using learning algorithms. In the case studied in this work, supervised learning algorithms were employed, which learn from labeled data where each input is associated with a corresponding target outcome. Additionally, among all the supervised learning algorithms, only regression algorithms were utilized, as there was a need to predict a numeric outcome. Within the realm of regression algorithms, both ensemble and non-ensemble models have been considered. The main difference between the two is that an ensemble model combines the predictions from multiple individual models to improve overall performance. Instead of relying on a single model, ensemble methods leverage the collective wisdom of multiple models to make more accurate predictions.
The final step in the machine learning process is result validation: once the model with the best performance on training data is obtained, it is assessed on new data, namely the test set. The concept of best performance and the parameters used to evaluate it are analyzed in the subsequent paragraph.

RESULTS EVALUATION

The analysis of the effectiveness of the regression model is carried out by assessing the values of one or more evaluation metrics chosen by the operator[4]. In this study, the chosen metrics to evaluate a single model and compare the predictive abilities of different models are the R^2 parameter and the mean absolute error (MAE) parameter.
R^2 Indicates how well the independent variables in a regression model explain the variability in the dependent variable. A key concept related to R^2 is variance, which is a measure of dispersion or variability of the data in a set of observations. In other words, variance indicates how much the values in a set of data deviate from their mean. The greater the variance, the greater the dispersion of the data around the mean; conversely, the smaller the variance, the smaller the dispersion of the data.
Mathematically, the formula for calculating the variance is:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The mathematical formula to calculate R^2 is:

$$1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_{true} - y_{pred})^2}{\sum(y_{true} - \bar{y})^2}$$

Where SS_{res} is the residual variance (the differences between the observed values of the dependent variable and the values predicted by the model) and SS_{tot} is the total variance, which represents the sum of squares of the

differences between the observed values of the dependent variable and the mean of the observed values of the dependent variable.
In practice, R^2 is calculated as the proportion of total variation in the dependent variable that is explained by the regression (i.e., how much the regression reduces the sum of squared residuals compared to the total sum of squares).
The value of R^2 ranges from 0 to 1 and can be interpreted as follows:

- R^2 =0 means that the model explains no variation in the dependent variable.
- R^2 =1 means that the model perfectly explains the variation in the dependent variable.

The closer R^2 is to 1, the better the model fits the data.
A low R^2 value could indicate that the regression model is not suitable for explaining the relationship between the variables or that important predictors are missing from the model.
However, a high R^2 does not necessarily imply that the model is predictive or that the correct model has been chosen.
In practice, R^2 is often used along with other evaluation metrics to assess the effectiveness of the regression model and determine if modifications or improvements are needed.
For this work, the second metric considered has been Mean Absolute Error (MAE), calculated by taking the average of the absolute differences between the predicted values and the actual values.
Its mathematical formula is:

$$\frac{1}{n}\sum|y_{true} - y_{pred}|$$

In simpler terms, MAE measures the average magnitude of the errors between the predicted and actual values. It gives an indication of how close the predictions are to the actual values on average. Additionally, being in the same unit of measurement as the data, it allows for a direct assessment of how significant the prediction error is relative to the data to be predicted.
Compared to other evaluation metrics, MAE is less sensitive to outlier because it considers absolute differences and extreme values have a linear effect on the error metric, thus the impact of outliers is not amplified.
For an overall evaluation of the model, is important to keep in mind that the primary objective of any machine learning model is to derive insights from data and generalize knowledge relevant to the task we are training it to perform.
Two main problems could arise during machine learning process:

- BIAS is the error introduced by approximating a model to the real relationships within the dataset. A model with high bias tends to be too simple and fails to capture the complexity of the data, leading to inaccurate predictions on both training and test data.
- VARIANCE measures how much the model's predictions vary for different sets of training data. A model with high variance tends to overfit noise in the data, resulting in very different predictions across different data sets and a poor predictive capacity on test data. An overfitting index is a strong deterioration between the metrics obtained in training, which have values very close to 1, and those obtained in testing.

In other terms, BIAS is related to underestimating relationships in the data, while VARIANCE is related to the model being overly sensitive to fluctuations in the training data. A good machine learning model seeks to balance bias and variance to achieve accurate and generalizable predictions[5].

How much data is enough?

Since machine learning relies on data, an obvious question is how much data is needed to train a model. In general, machine learning performs best when the training set includes a comprehensive range of feature combinations. In other words, the more diverse the combinations available in the dataset, the more effective the model will be at capturing the impact of each feature on the dependent variables. However, as a rule of thumb, the absolute minimum amount of data required is ten times the total number of independent variables or features.
While data is essential to the self-learning process, simply having more data doesn't always lead to better decisions; it's the relevance of the input data that really matters. Adding irrelevant data is counterproductive, firstly because it could obscure the pattern being identified, and secondly because more data to analyze means more time and processing resources needed for the analysis.

DEVELOPMENT ENVIRONMENT

To fulfill the objectives of this study, we chose to utilize the tools provided by Azure Machine Learning Studio, a cloud platform by Microsoft tailored for developing and training machine learning models[6]. Initially, we conducted some tests using the open-source Python library "Scikit Learn" in a Linux environment. However, the need for more extensive computational resources than those available prompted us to transition the entire study to a cloud-based environment.
Azure Machine Learning Studio offers the capability to create multiple computation clusters with dedicated resources, allowing the simultaneous launch of multiple model trainings. Additionally, it provides robust Auto Machine Learning tools to automatically identify the most effective algorithms.
The platform boasts a user-friendly graphical interface, enabling users to initiate trainings without writing a single line of code. Despite its intuitiveness, it's crucial to carefully configure training settings to avoid misinterpreting output data or overlooking issues like overfitting or suboptimal hyperparameter settings.
A critical consideration is the proper configuration of processing clusters; as a Microsoft cloud service, it entails costs, necessitating resource allocation according to actual requirements to avoid unnecessary expenses.
In this study, the Auto Machine Learning tool was primarily utilized to determine the best-suited algorithm for regression analysis. To effectively utilize this tool, four fundamental steps were followed:

- Creating datasets compatible with the cloud environment
- Importing datasets
- Configuring training parameters
- Evaluating the results

Regarding dataset creation, it's important to note that files must be in .csv format and, for numerical data, the decimal separator must be a "." to avoid data reading and conversion errors. After obtaining the correct format, datasets were imported, and necessary adjustments, such as defining data types for each column, were made before importation.
Subsequently, training settings were configured, specifying the type of training (e.g., regression, classification), selecting the target column, and setting parameters like evaluation metrics and the dataset split ratio (in this study, always 70% - 30%).
Upon completion of training, it's important to verify model output data to assess its effectiveness and identify areas for improvement.
An interesting feature of this tool is its ability to register and publish models as real-time endpoints accessible from any web app via address and key provided by the service.

RESULTS

FROM COMPOSITION TO HARDNESS AS ANNEALED

The first physical characteristic investigated was the hardness after cold working (75%) and annealing. To better understand the influence of the number of data points on the final results, several trainings were conducted, starting from a limited number of initial data points and gradually expanding to the full dataset.
To achieve this, the alloys with post-annealing hardness data were divided into different categories based on karatage and color (Table 1). An equal number of elements was chosen for each category to ensure that, even after data reduction, the initial distribution was maintained, thereby preserving the maximum amount of information.

Table 1: Division into categories of characterized alloys

	RED GOLD ALLOYS	YELLOW GOLD ALLOYS	WHITE GOLD ALLOYS	WHITE GOLD ALLOYS WITH PD	SILVER ALLOYS
(375)	18	14	9		
(417)	17	12	9		
(585)	23	33	16	1	
(750)	23	24	28	17	
(875)		6			
(917)		4			
(925)					19

The prediction results for each dataset are shown in Table 2: it is evident that as the number of data increases, there is a corresponding improvement in evaluation metrics. Regarding the training set, the R^2 value increases from 0.66 to 0.84, while the mean error decreases from 16 HV to 11 HV. In the case of the test set, the improvement is even more pronounced: for the first two datasets, R^2 is actually negative, a value obtained when the residual deviation is greater than the absolute deviation, meaning that the error generated by the model on unknown data is even greater than the error generated by a potential model that uses the data average as output. As for the MAE of the test, we achieve a minimum value of 9 HV on the complete dataset, a good predictive result considering that the experimental error is usually ± 15 HV.

Table 2: training parameters and results for hardness as annealed

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
68 + 28	19	0.66	16	-0.47	38
121 + 51	19	0.75	12	-0.67	38
161 + 70	19	0.85	11	0.82	11
193 + 84	19	0.84	11	0.89	9

FROM COMPOSITION TO HARDNESS AFTER AGE HARDENING

Maximum hardness after age hardening was similarly utilized to develop a predictive model. In Table 3, the number of alloys with hardness values after age hardening is indicated, divided for each category. The initial training conducted on this physical characteristic encompassed the entire dataset available.

Table 3: Hardenable alloys and their distribution into categories.

	RED GOLD ALLOYS	YELLOW GOLD ALLOYS	WHITE GOLD ALLOYS	WHITE GOLD ALLOYS WITH PD	SILVER ALLOYS
(375)	5		14	3	
(417)	3		11	4	
(585)	6	3	27	3	
(750)	23	30	36	20	
(875)			3		
(917)			1		
(925)					29

In Table 4, the total number of data points (divided into training and testing sets), the number of features and the training results are listed together with the R^2 and MAE metric values on both the training and testing sets.

Table 4:Training parameters and prediction evaluation for hardness after age hardening with the complete dataset.

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
157 + 68	19	0.82	17	0.87	16

From the evaluation parameters analysis, it is evident that the R^2 value of the training set is similar to that of the testing set, suggesting a good model generalization capability. In Figure 1, the individual prediction errors for each alloy of the test set are visualized in detail, sorted by increasing hardness value.

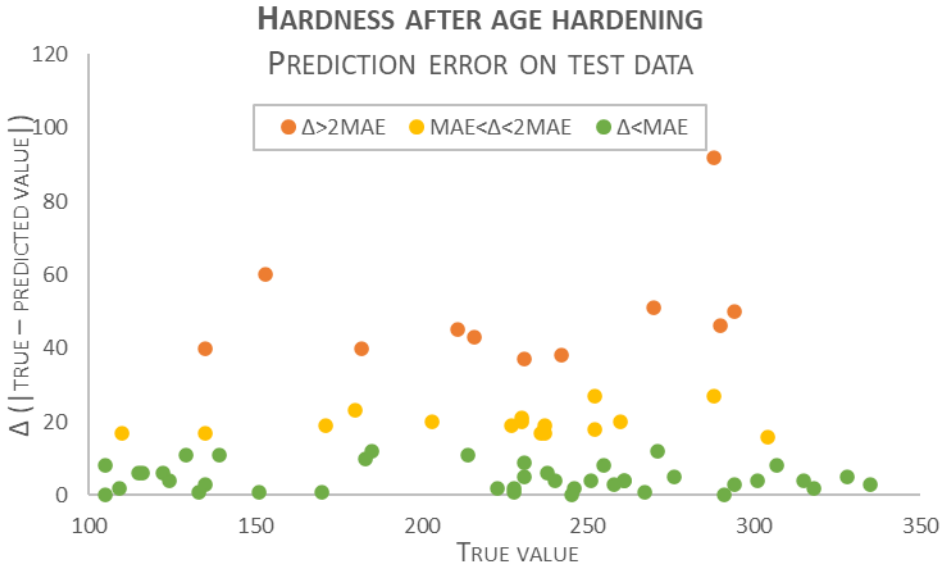

*Figure 1: **Prediction error in the test dataset for hardness after age hardening, full dataset.***

Comparing the evaluation parameter values with the results obtained for hardness after annealing, it can be observed that the R^2 value (for both the training and testing sets) is comparable, while the mean absolute error is worse in both cases. The R^2 value of 0.87, obtained from the testing set for predicting hardness after annealing and after age hardening, indicates that in both cases, approximately 87% of the variance in the dependent variable can be explained by the independent variables.
Although overall the predictive capability of the model is comparable to the case of hardness after annealing, considering the R^2 values, we must expect a higher error in predicting the hardness value after hardening compared to that of hardness after annealing (16 HV versus 9 HV) using this trained model. In an attempt to reduce the mean prediction error, further experiments were conducted by varying the training conditions. For subsequent training, the dataset was narrowed down to evaluate the predictive capacity within the most characterized region of the alloy space. Table 5 shows the total number of alloys for each karatage; it was decided to limit the dataset to 18 karats, which is the most characterized karatage with 109 alloys (Table 5).

Table 5: Sum for each karatage of the alloys with age hardening data

	RED GOLD ALLOYS	YELLOW GOLD ALLOYS	WHITE GOLD ALLOYS	WHITE GOLD ALLOYS WITH PD	SILVER ALLOYS	SUM
(375)	5		14	3		22
(417)	3		11	4		18
(585)	6	3	27	3		39
(750)	23	30	36	20		109
(875)			3			3
(917)			1			1
(925)					29	29

Table 6:Training parameters and prediction evaluation for hardness after age hardening, only 18k

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
76 + 33	19	0.66	15	0.72	16

The result, reported in Table 6 and Figure 2, highlights a MAE very similar to the previous case, but a worse R^2. Even though the selected region in the alloy composition space is densely populated in terms of characterized combinations, the effect of an overall limited number of data points compared to the number of features (109 data points for 19 elements) impacts the predictive outcome, making it overall worse than the previous case.

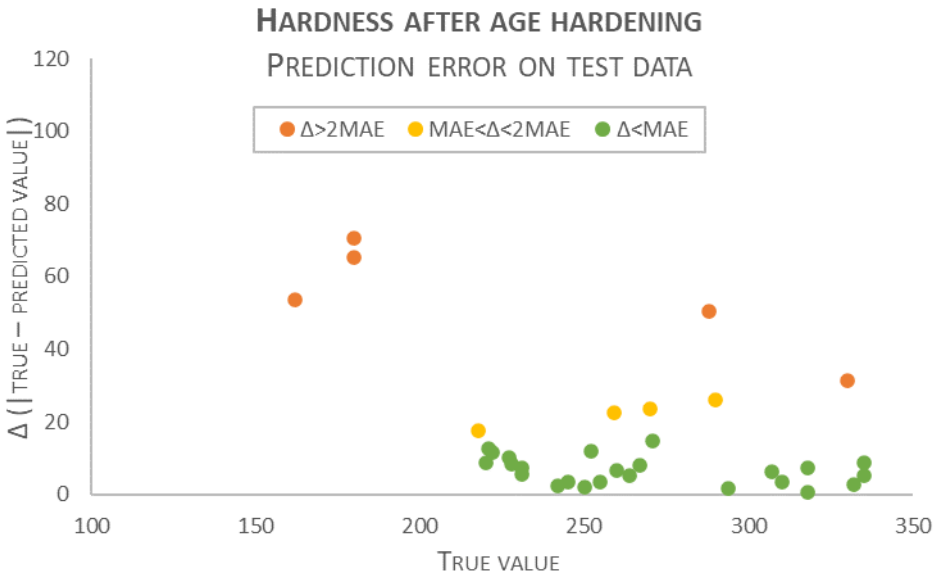

Figure 2: Prediction error in the test dataset for hardness after hardening, 18 k only.

In the subsequent training trial, a predetermined split of the training and test sets was chosen instead of relying on the random division of the initial dataset. This predefined split was specifically designed to ensure a balanced representation of all analyzed alloy categories in both sets, maintaining the 70% data ratio in the training set and 30% in the test set.

Table 7: Training parameters and prediction evaluation for hardness after age hardening, full dataset, predefined train and test set.

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
157 + 68	19	0.82	16	0.81	18

The results, as shown in Table 7 and Figure 3, indicate a predictive capacity in the training set comparable to the first learning trial but lower in the test set. The lack of improvement in the evaluation parameters suggests that in the first learning trial, the random division of the initial dataset performed by the program already ensured a homogeneous representation of the various alloy categories between the training and test sets.

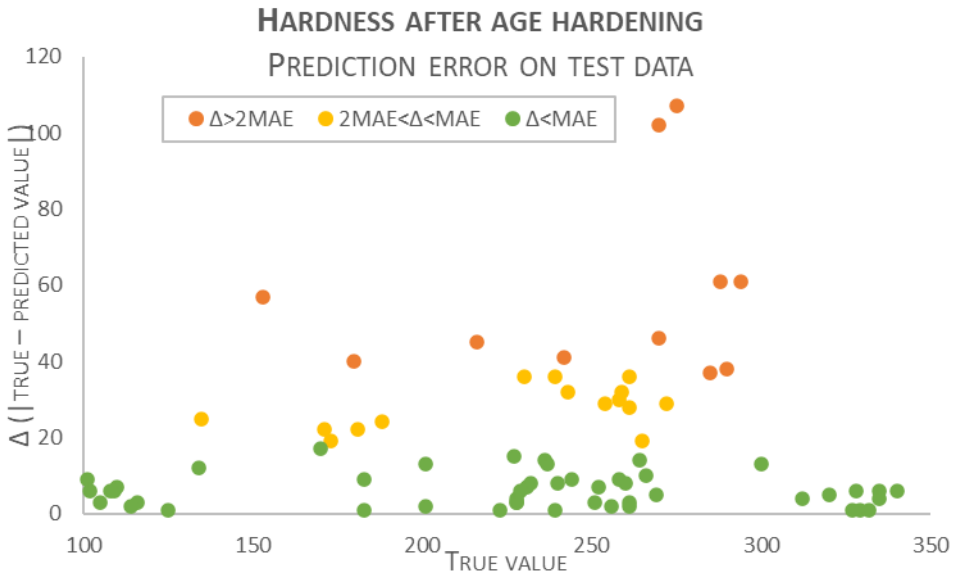

Figure 3: Prediction error in the test dataset for hardness after hardening, full dataset, predefined train and test set

Finally, an experiment was conducted to reduce the input features through a semi-empirical approach. Feature reduction, commonly known as dimensionality reduction, is a widely adopted process in machine learning to decrease the number of input variables or features in a dataset. The main objective is to simplify the dataset while retaining its essential characteristics, which can contribute to improving the performance of machine learning models. The advantages of this approach mainly consist of:

- Improved model performance: By eliminating irrelevant features, models can focus on the most informative aspects of the data, often achieving better performance.
- Reduction of overfitting: A smaller number of features results in lower complexity, which can reduce the risk of overfitting.

Generally, this process is performed using computational algorithms, but in this case, a more empirical approach was chosen by combining machine learning with experimental knowledge. From the list of input variables (i.e., elements in the composition), those that, according to experience, do not have a significant impact on the maximum hardness value were eliminated. Additionally, elements that contribute to the final hardness value but are represented very minimally in the sample (e.g., only 2 or 3 alloys in total have this element in their composition) were not considered, and alloys with these elements were removed from the dataset.

Table 8: Training parameters and prediction evaluation for hardness after age hardening, full dataset, reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
157 + 68	15	0.79	18	0.91	13

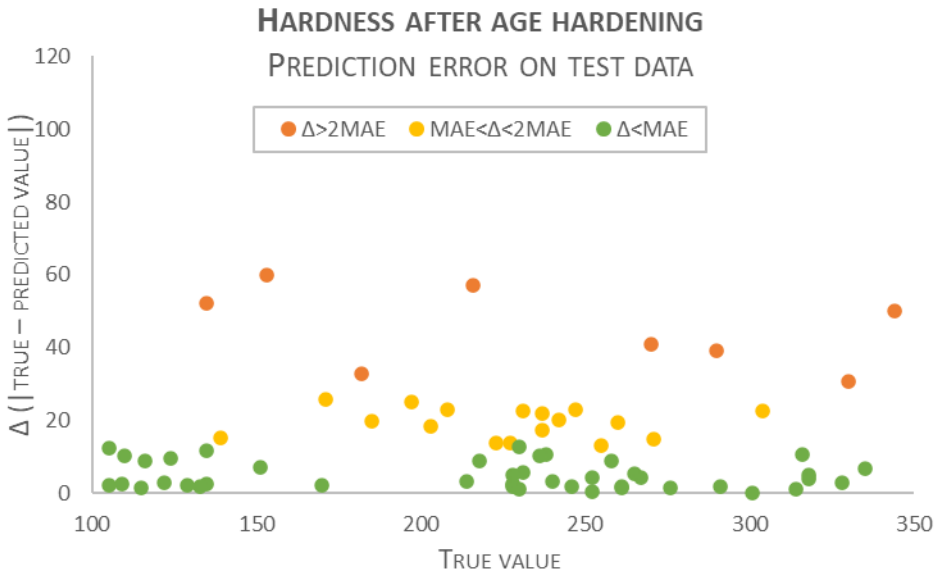

Figure 4:Prediction error in the test dataset for hardness after hardening, full dataset, reduced features.

This training resulted in the best predictive performance on the test set (Table 8 and Figure 4), explaining 91% of the variance. However, compared to previous cases, there is a significantly better performance on the test set than on the training set. One of the main, but not necessarily sole, reasons for this discrepancy could be the presence of an outlier in the training set. For example, it could be an alloy with unusual behavior, perhaps because it is the only one in its composition range or due to an error in the recorded experimental data.
To verify this hypothesis, the trend of errors in the training set was analyzed in search of anomalous results, and indeed, an alloy with a prediction error of 27 HV was found, significantly higher than any other error in the set. Upon observing the composition and hardness value, the likely explanation was found: it is the only 14 karats gold alloy in the entire dataset that does not harden due to its low percentage of silver. The hardness value after hardening, which is normally left blank for non-hardening alloys, had nevertheless been included in the alloy's characterization data and therefore used for training. It is worth noting that in previous learnings, the impact of this alloy on the evaluation parameter values was not as pronounced (it was not even considered in Learning 2 as it is a 14 kt alloy), as the prediction error on the outlier in these cases was not significantly higher compared to some of the prediction errors on actually hardening alloys.
The next step was therefore to remove this alloy from the dataset and redo the training.

Table 9: Training parameters and prediction evaluation for hardness after age hardening, full dataset, reduced features and outlier removal

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
156 + 68	15	0.85	17	0.93	11

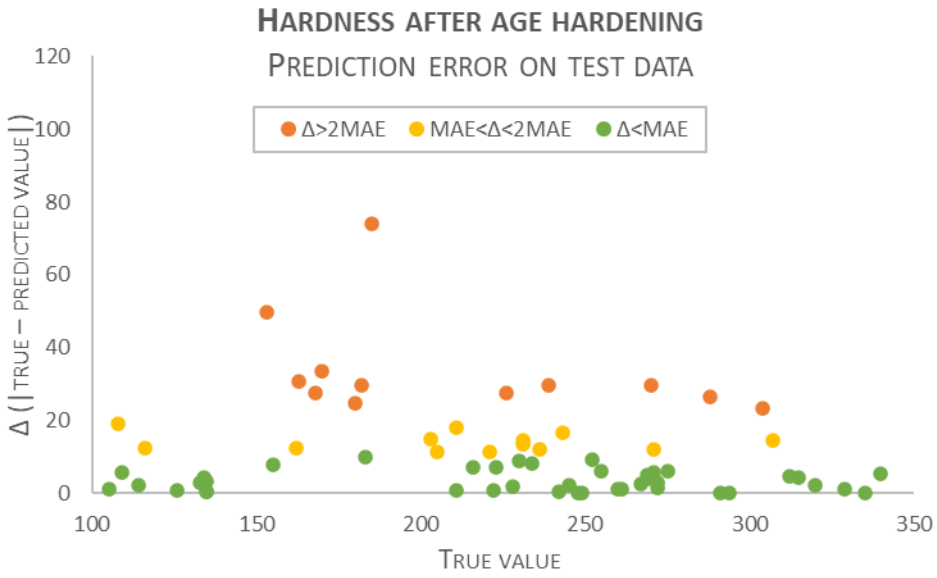

Figure 5: Prediction error in the test dataset for hardness after hardening, full dataset, reduced features and outlier removal.

The results (Table 9 and Figure 5) show an improvement in the evaluation parameters both in the training and testing sets, but again, better values are observed for the test set compared to the training set. Since the presence of other outliers has been excluded, the explanation may lie in the small size of the entire dataset and consequently of the test set, which represents only 30% of it: it is possible that the test set does not include some particular cases that are present in the training set and on which the model does not perform optimally.

FROM COMPOSITION TO COLOR

Coordinate L*

Similarly to what was done with hardness, the first training for predicting the L* color coordinate also involved the entire dataset (results in Table 10), without reducing the features. The experimental data on which the training is based were taken with D65 illuminant and wide observation angles (10°).

Table 10: Training parameters and prediction evaluation for L coordinate, full dataset*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
277 + 120	19	0.90	0.60	0.92	0.66

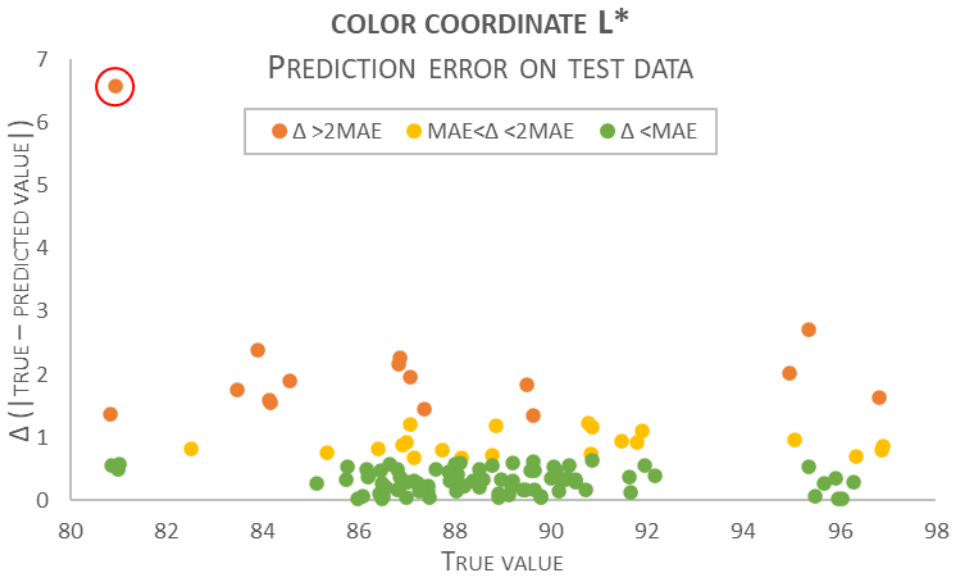

Figure 6: Prediction error in the test dataset for color coordinate L, full dataset.*

Observing the trend of errors in the test set (Figure 6), it is evident the presence of an outlier with a very high error. Analyzing the L* value of the alloy in question and comparing it to the measured value on alloys of similar composition, a probable experimental error in the measurement or transcription of the data has been noticed. The color of the alloy was then remeasured, obtaining an actual L* value of 85.61 instead of 80.95, much closer to the predicted value. This confirms that in this case, the error in the test was not caused by a poor prediction but rather by an incorrect starting data.
The training was then repeated, correcting the erroneous data (Table 11). The evaluation parameter values show an improvement in both the training and test sets. Observing the error values of individual test data points (Figure 7), a couple of points with significant errors are still noticeable, whose corresponding L* values, however, are not affected by experimental errors: in this case, they truly represent incorrect predictions.

Table 11: Training parameters and prediction evaluation for L coordinate, full dataset and outlier correction*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
277 + 120	19	0.92	0.57	0.94	0.58

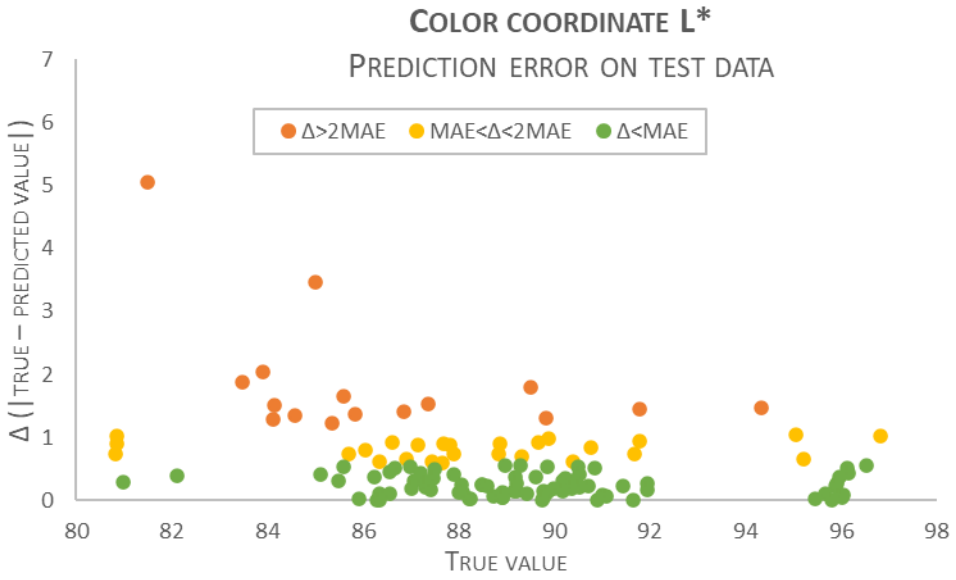

Figure 7: Prediction error in the test dataset for color coordinate L, full dataset and data correction.*

The second training on the L* coordinate focused solely on Au alloys, across all karatage. Once again, as with post-hardening hardness, the aim was to concentrate on compositions with a higher density of data points, while avoiding overly restricting the number of points considered compared to the possible alloy constituents. The result highlights that, even in this case, there is no actual improvement in predictive capability. Observing the error trends for the test set (Figure 8), a single point with a higher error is noticeable, which, however, does not correspond to problematic alloys in the first training and again is not due to experimental errors.

Table 12: Training parameters and prediction evaluation for L coordinate, gold alloys only*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
256 + 110	19	0.90	0.57	0.83	0.60

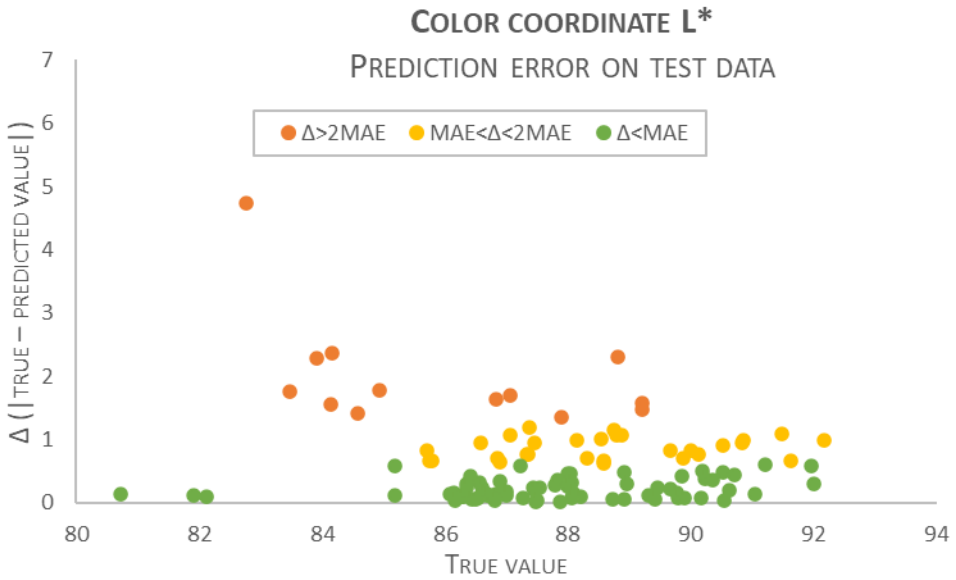

Figure 8: Prediction error in the test dataset for color coordinate L, gold alloy only.*

The final training was conducted on the entire dataset (including Au and Ag alloys), but with a reduction in features. This reduction excluded elements that do not contribute to the overall color of the alloy or are present in a very limited number of samples. Similarly to hardness, this process was carried out empirically, based on

experimental knowledge. It's worth noting that the reduction of elements applied to color coordinates was greater than that for hardness, resulting in 12 remaining features compared to 15.

Table 13: Training parameters and prediction evaluation for L coordinate, full dataset and reduced features*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
278 + 120	12	0.94	0.52	0.93	0.61

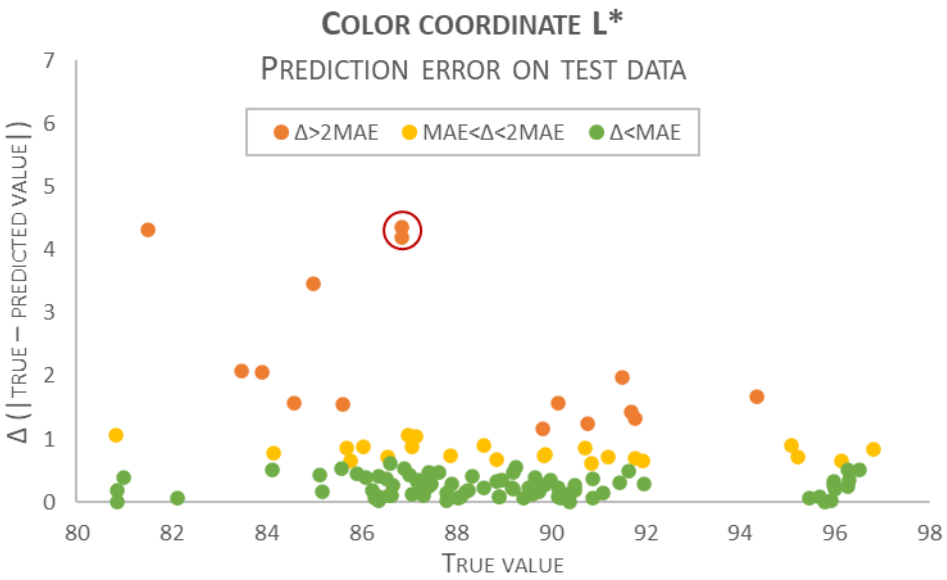

Figure 9: Prediction error in the test dataset for color coordinate L, full dataset and reduced features.*

The results (Table 13) highlight an improvement for the training dataset, but this improvement is not accompanied by a similar enhancement in the test set. Observing the distribution of errors (Figure 9) emphasizes three points with errors exceeding 4, which were then analyzed more thoroughly. For the two points with an L* value close to 86.8, a probable explanation was found: they are the only two 18k Au-Pd alloys that contain a certain element in their composition. Since both alloys are in the test set for this model training, this combination of elements is not present in the training set. Therefore, the model is not prepared to correctly predict their brightness.
To address this issue, further training was conducted with predefined training and test sets identical to the previous case, with only one of the two alloys moved from the test set to the training set (Table 14 and Figure 10).

Table 14: Training parameters and prediction evaluation for L coordinate, full dataset, reduced features and predefined train and test set*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
278 + 120	12	0.92	0.55	0.94	0.57

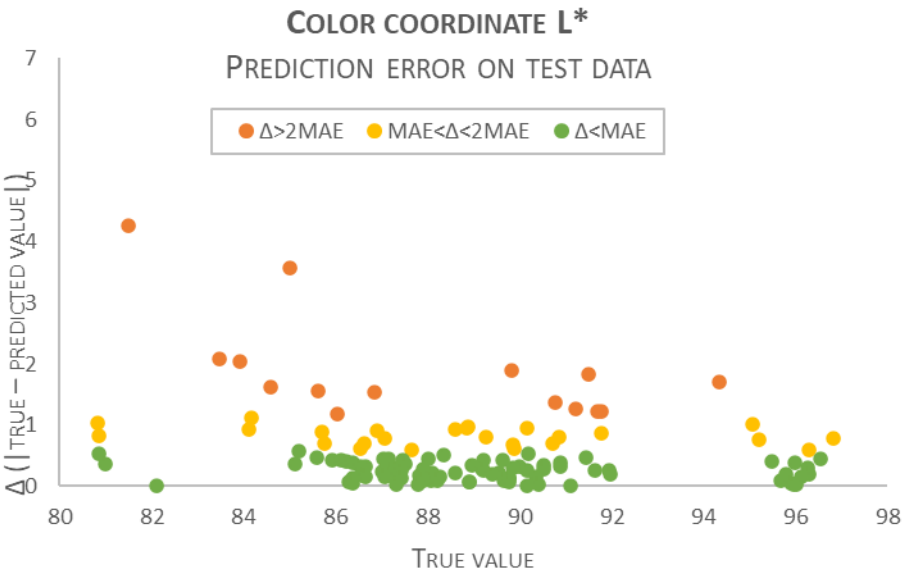

Figure 10: Prediction error in the test dataset for color coordinate L, full dataset, reduced features and predefined train and test set.*

Coordinate a*

As with the L* coordinate, for the a* coordinate (and subsequently for the b* coordinate), training trials first involved the complete dataset, then only Au alloys, and finally the complete dataset with feature reduction. Regarding the complete dataset, it is visible in Table 15 that the predictive behavior is overall better compared to the case of the L* coordinate, both in terms of R^2 and MAE. This fact also reflects the trend of experimental brightness measurements, which depend greatly on the quality of polishing and therefore the operator's skill. In addition, there are no points with a prediction error that stand out from the rest (

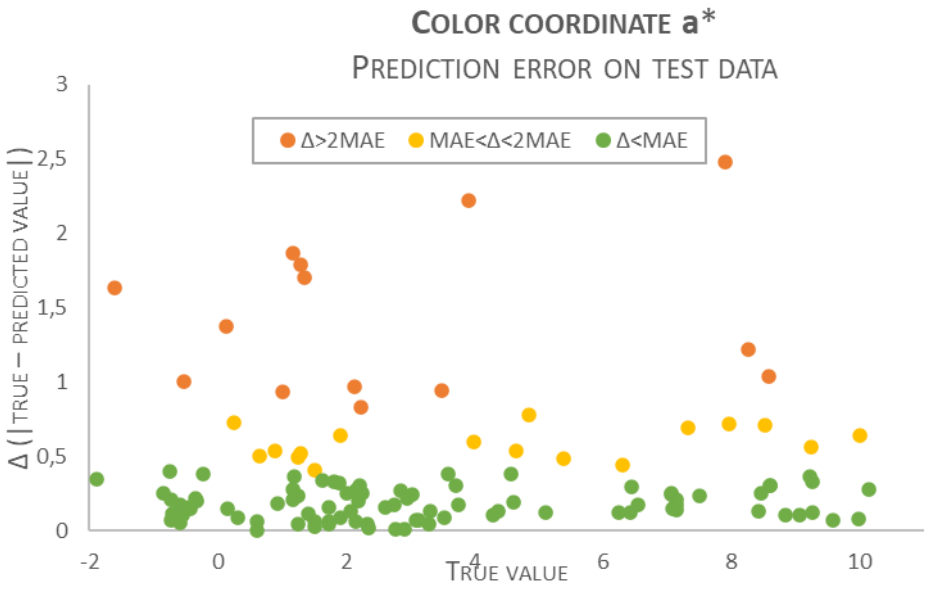

Figure 11).

Table 15: Training parameters and prediction evaluation for color coordinate a, full dataset.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
278 + 120	19	0.94	0.52	0.96	0.4

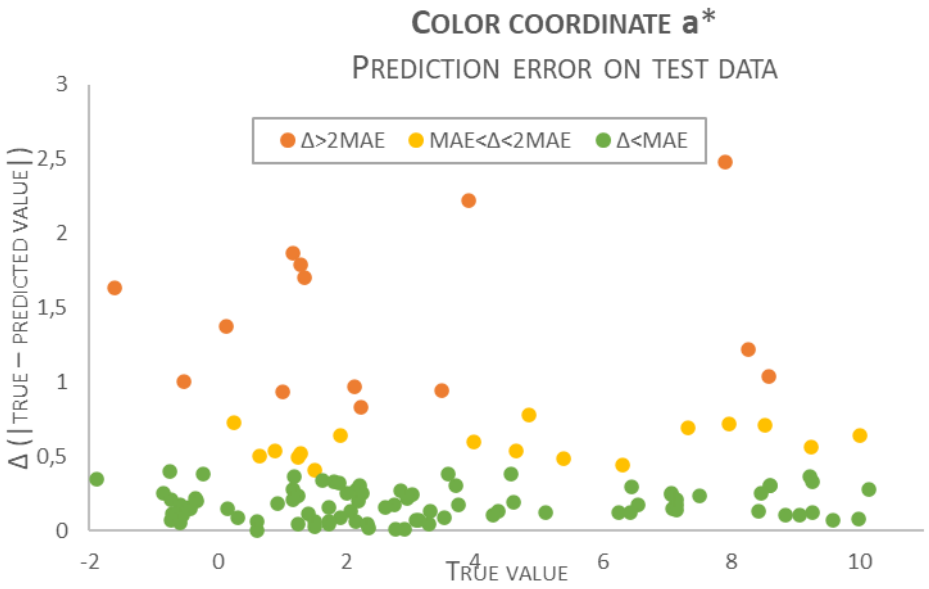

Figure 11: Prediction error in the test dataset for color coordinate a, full dataset.*

By limiting the dataset to gold alloys only, the validation parameter values are more consistent between the training and test sets (Table 16). However, overall, there is no improvement observed in the predictive capability of the model. Once again, there are no outliers with abnormal prediction errors (

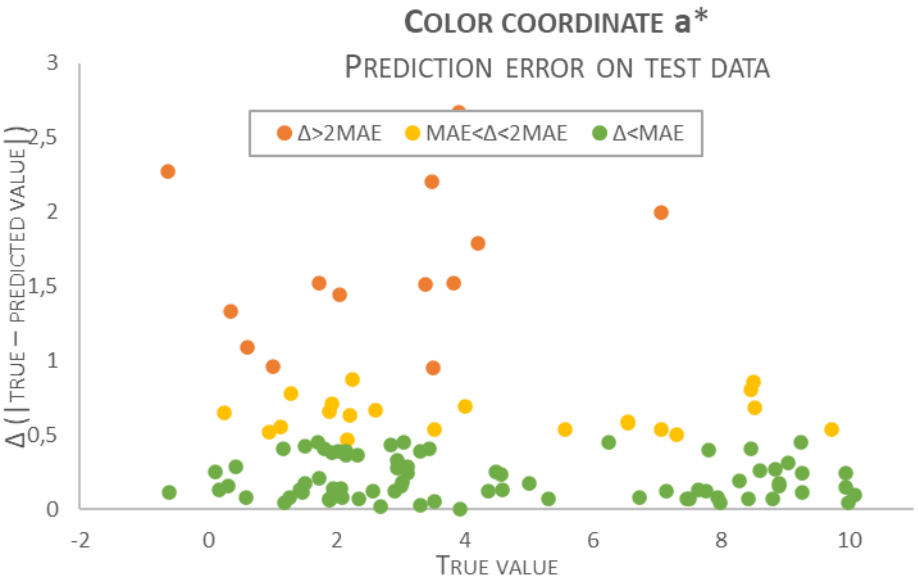

Figure 12).

Table 16: Training parameters and prediction evaluation for color coordinate a, gold alloy only*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
256 + 110	19	0.95	0.46	0.95	0.45

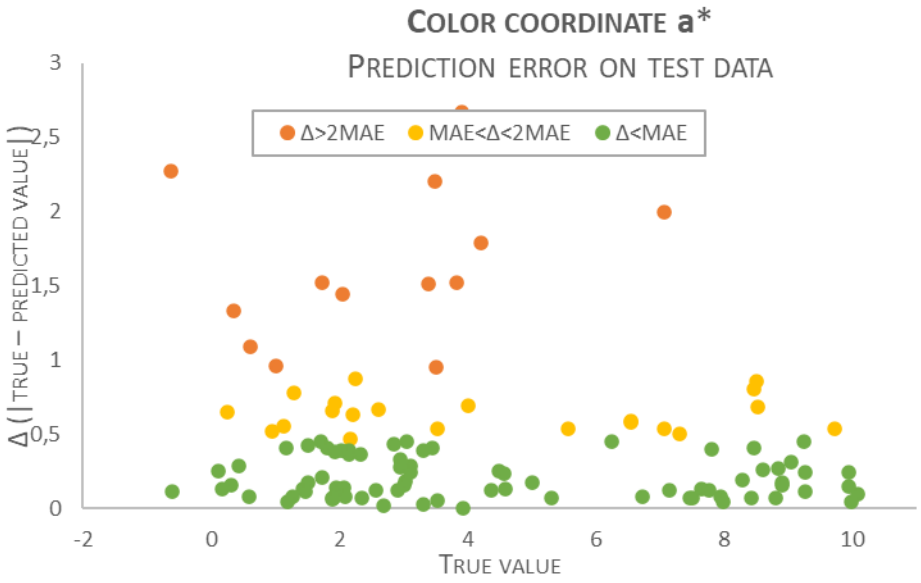

Figure 12: Prediction error in the test dataset for color coordinate a, gold alloy only.*

Evaluating the training performed on the entire dataset but with reduced features (Table 17), it's noticed that the results for the training set are similar to previous trainings, while there is a slight improvement in the test phase. The difference between training and test, with the latter being more performant again, led to a check for any outliers in the training set, but none were identified.

Table 17: Training parameters and prediction evaluation for color coordinate a, full dataset and reduced features.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
277 + 120	12	0.94	0.5	0.97	0.34

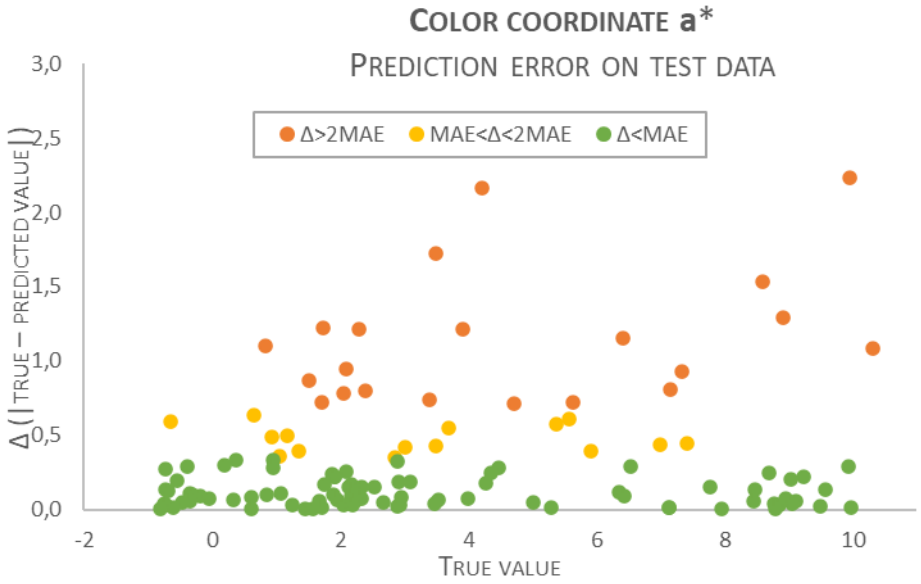

Figure 13: Prediction error in the test dataset for color coordinate a, full dataset, reduced features.*

Coordinate b*

Training on the complete data set of coordinate b* yielded intermediate metric values between those obtained for L* and for a* under the same conditions (Table 18).

Table 18: Training parameters and prediction evaluation for color coordinate b, full dataset.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
277 + 120	19	0.94	0.6	0.98	0.53

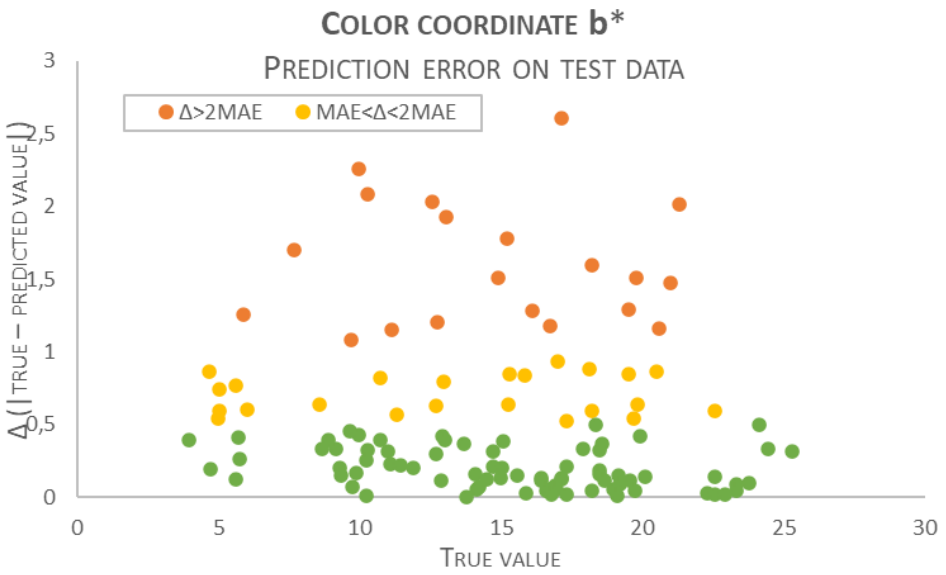

Figure 14: Prediction error in the test dataset for color coordinate b, full dataset*

If the training is performed only on the dataset of gold alloys, it's immediately apparent an outlier with a very high error in the test set (Table 17 and Figure 15)

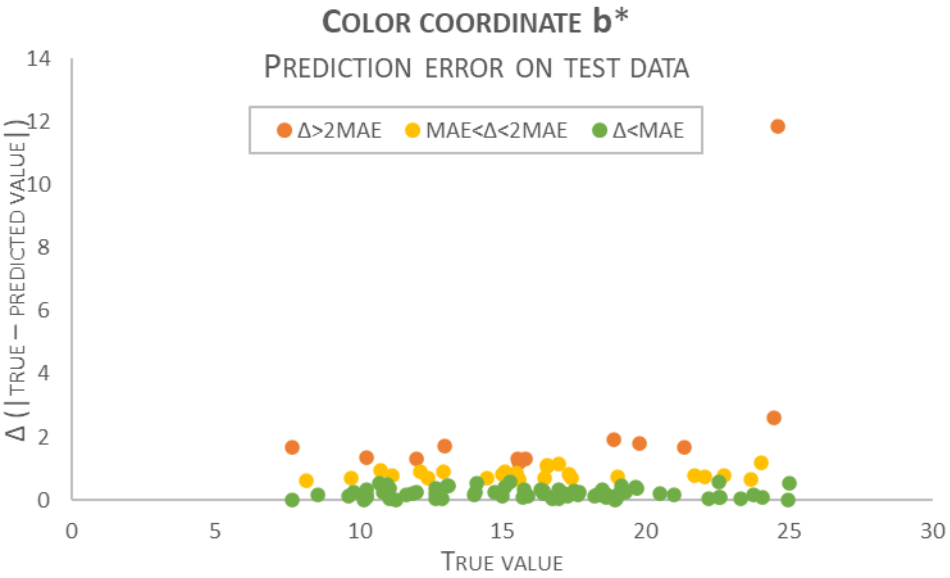

Table 19: Training parameters and prediction evaluation for color coordinate b, gold alloys only.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
256 + 110	19	0.95	0.57	0.91	0.60

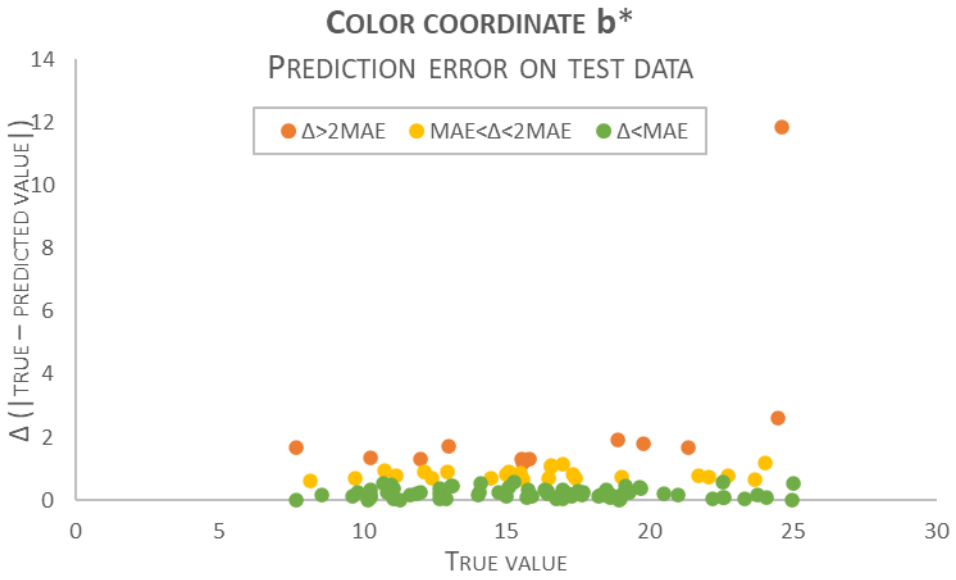

Figure 15: Prediction error in the test dataset for color coordinate b, full dataset, gold alloys only*

An in-depth analysis of the actual and predicted values of the b* coordinate of the corresponding composition revealed, once again, a likely error in the recorded experimental data. Repeating the measurement confirmed this suspicion, with a measured value of 11.22 compared to the previous 24.6. For the first training on b*, the alloy with the erroneous data was in the training set, not in the test set, thus was not visible from Figure 14. The error in predicting the b* coordinate however was in that case not significantly different from that recorded for some of the other alloys, making it less visible and contributing less to the R^2 and MAE values.
Training on gold alloys only was subsequently repeated with the corrected data (Table 20 and Figure 16), resulting in better prediction results because they were not distorted by the error in the b* coordinate value of the alloy under examination.

Table 20: Training parameters and prediction evaluation for color coordinate b, gold alloys only and outlier correction.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
256 + 110	19	0.96	0.50	0.97	0.55

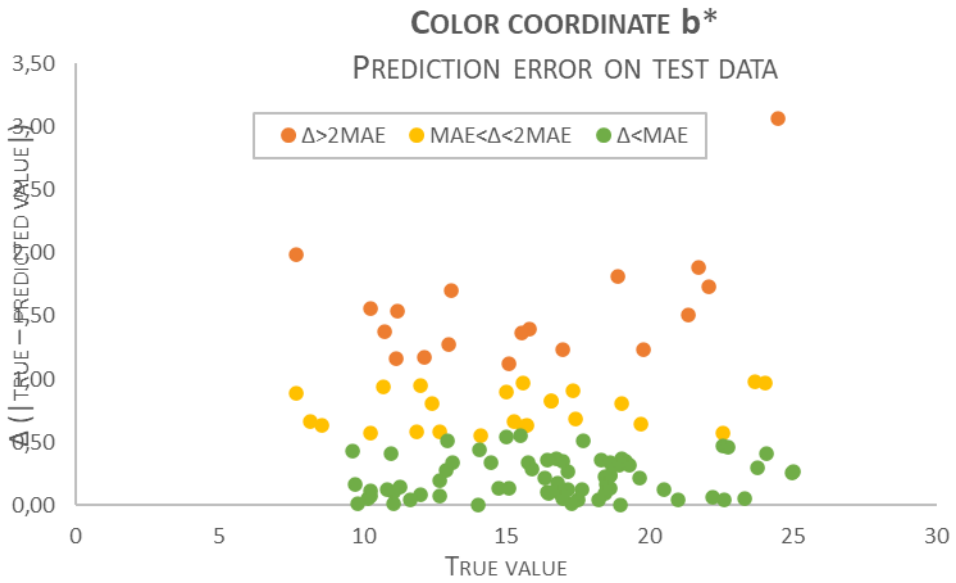

Figure 16: Prediction error in the test dataset for color coordinate b, gold alloys only and outlier correction.*

By reducing the number of features, as previously done for coordinates L* and a*, the evaluation parameter values remain similar for the training set to those seen with gold alloys, while improving in the test set compared to previous trainings (Table 21). This is despite two points (clearly visible in Figure 17) having an error that exceeds the maximum error obtained with the previous trainings, outliers aside. In all previous trainings, these two alloys were in fact in the train set, not the test set, and their predictive error is therefore not present in Figure 14, Figure 15 and Figure 16. The re-measurement of b* values for these two alloys did not reveal any previous experimental errors, thus leaving the probable cause of the high prediction errors to the compositions of the alloys themselves, which are part of a poorly characterized zone in the space of possible alloys.

Table 21: Training parameters and prediction evaluation for color coordinate b, full dataset and reduced features.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
277 + 120	12	0.96	0.52	0.97	0.43

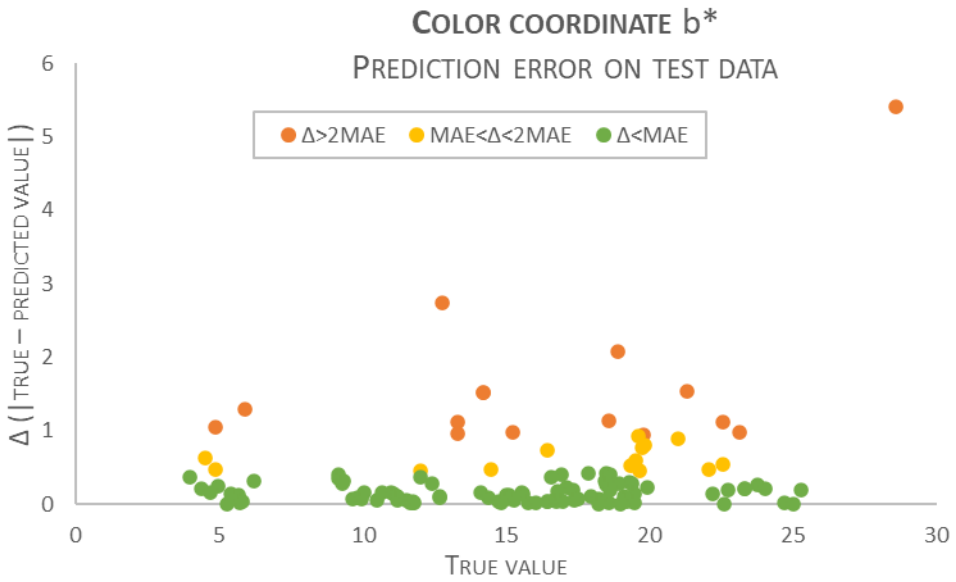

Figure 17: Prediction error in the test dataset for color coordinate b, full dataset and reduced features.*

COLOR COORDINATES CONVERSION

Machine learning has also been used to determine the change in CIELAB coordinates values when transitioning from an observer with a wide field of view (10°) to an observer with a narrow field of view (2°), always with D65 illuminant. In ISO 8654:2018, the observer with a narrow field of view is indeed reported as the primary one, while still maintaining an appendix regarding measurements with a wide field of view. In cases where replicating colorimetric measurements with a different field of view is not possible, conversion is usually feasible using traditional color calculation software if the complete spectrophotometric curve is available. However, direct conversion from simple L*, a*, and b* values is generally not allowed. Similarly, there are no analytical formulas that allow conversion of values from one measurement angle to another.
Having the need to convert some of the characterization data taken in the past, for which complete curves were not available, it was therefore considered to use artificial intelligence to train a model capable of predicting the coordinate values with the new viewing angle.

Coordinate L*

Similarly to previous cases, the first training was conducted for all coordinates using the entire available dataset. Additionally, it was chosen to take only the coordinate of interest as the independent variable and not all three. Thus, in this case, the value of L* with measurement D65/02° is predicted solely from the value of L* measured with D65/10°. The results are shown in Table 22: notably, there are significantly better values in the evaluation

parameters compared to those observed in predicting the L coordinate from the composition, and they are similar between the training and test sets.

Table 22: Training parameters and prediction evaluation for color coordinate L, full dataset.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	1	0.98	0.20	0.99	0.17

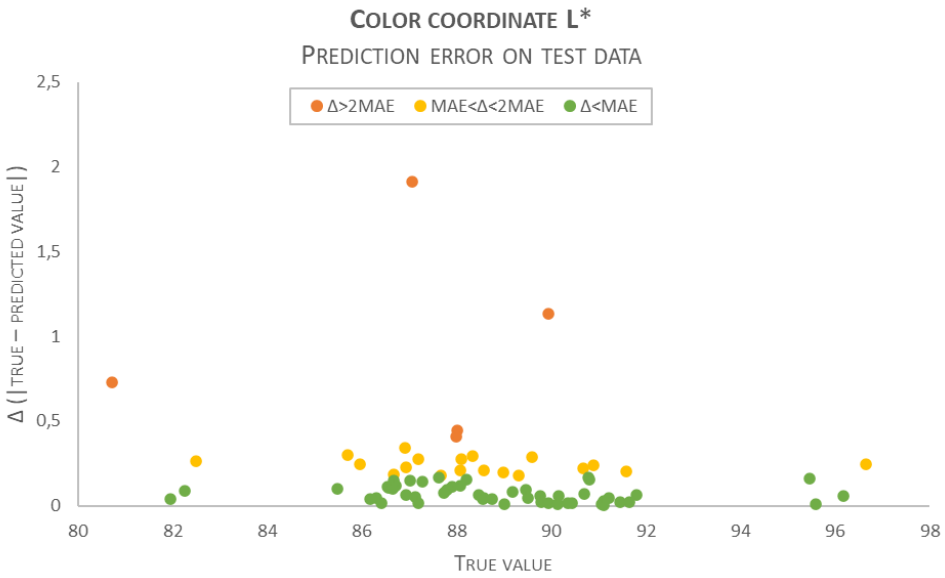

Figure 18: Prediction error in the test dataset for color coordinate L, full dataset*

In the second training, the value of L* measured with D65/2° was predicted considering all three coordinates L*, a*, b* (measured with D65/10°) as independent variables. The results (Table 23, Figure 19) highlight a further improvement in predictions, both in terms of the R^2 parameter and the mean absolute error.

Table 23: Training parameters and prediction evaluation for color coordinate L, full dataset and L,a,b D65/10° as dependent variables.*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	3	0.99	0.14	0.99	0.12

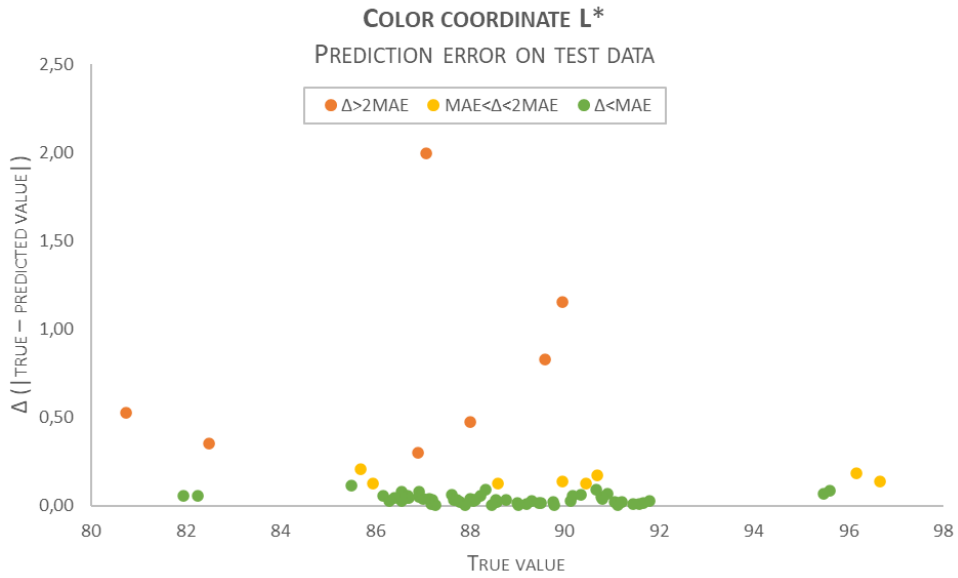

Figure 19: Prediction error in the test dataset for color coordinate b, full dataset and L,a*,b* D65/10° as dependent variables.*

Coordinate a*

As with the L* coordinate, the first training for the coordinate a* also occurred with the entire dataset, considering only the same coordinate under D65/10° conditions as the independent variable.
The results (Table 24) are significantly worse than those obtained with the L* coordinate. Observing the distribution of errors in the test set (Figure 20), there is a deterioration in predictive capability at the lower and upper extremes of the a* value range, i.e., between -2 and 0 and between the values 8 and 10.

Table 24: Training parameters and prediction evaluation for color coordinate a, full dataset*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	1	0.96	0.48	0.96	0.46

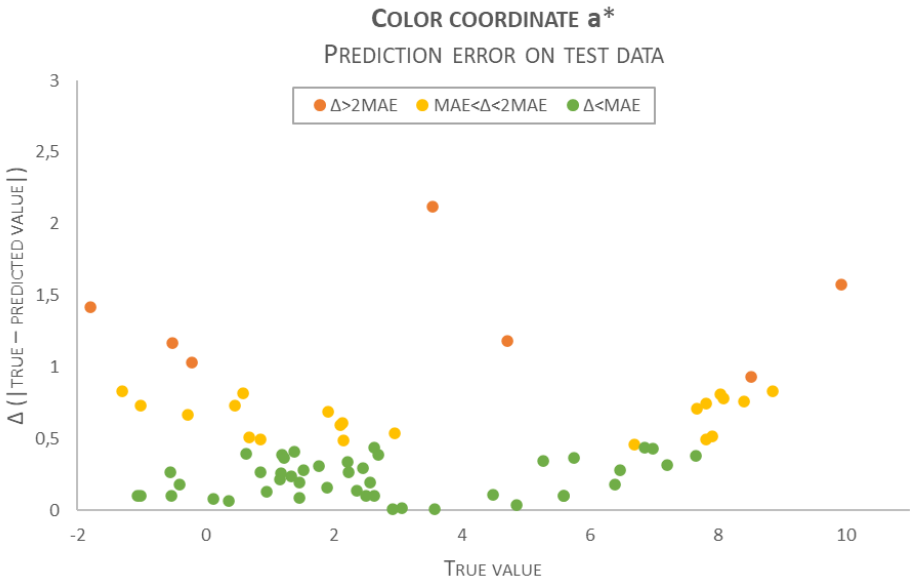

Figure 20: Prediction error in the test dataset for color coordinate a, full dataset.*

Including L* and b* among the independent variables results in a significant improvement in predictive capability (Table 25), with R^2 and MAE values practically identical to those obtained for the L* coordinate under the same training conditions. The error distribution also appears much more random (Figure 21).

Table 25: Training parameters and prediction evaluation for color coordinate a, full dataset and L*,a*,b* D65/10° as dependent variables*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	3	0.99	0.14	0.99	0.12

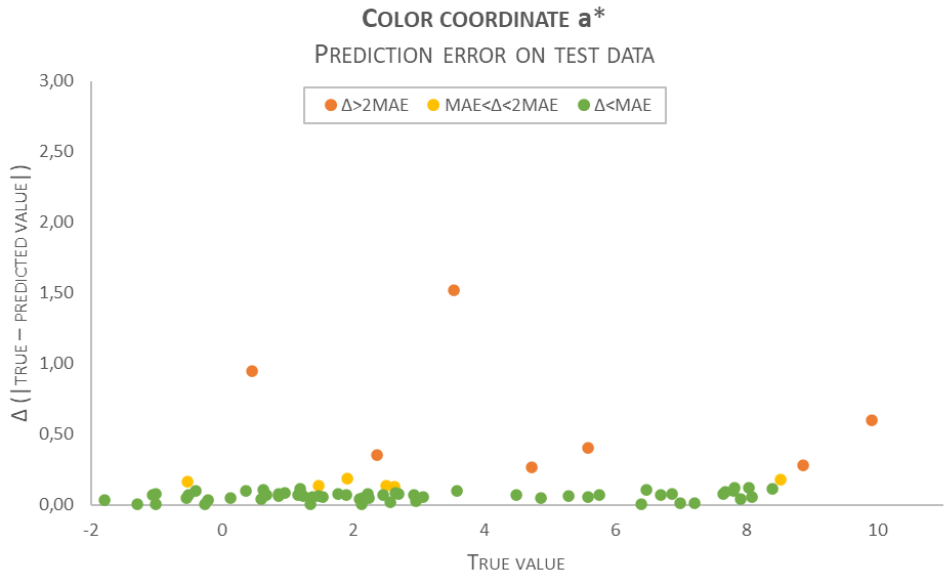

Figure 21: Prediction error in the test dataset for color coordinate L, full dataset and L*, a*, b* D65/10° as dependent variables.*

Coordinate b*

For the b* coordinate, a similar behavior to that of the L* coordinate was observed: even with only one independent variable (b* value with D65/10°), the evaluation parameters show good values (Table 26).

Table 26: Training parameters and prediction evaluation for color coordinate b, full dataset*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	1	0.99	0.26	0.99	0.23

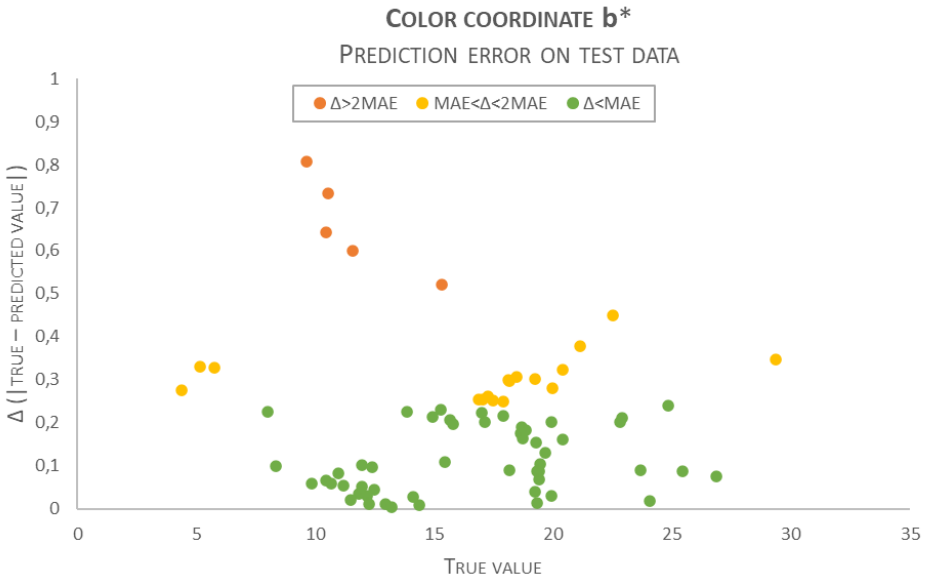

Figure 22: Prediction error in the test dataset for color coordinate b, full dataset*

By including L* and a* as variables, there is a further improvement (Table 25Table 27), as observed for the other two coordinates, although it is more limited in terms of mean absolute error.

Table 27: Training parameters and prediction evaluation for color coordinate b, full dataset and L*, a*, b* D65/10° as dependent variables*

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
170 + 74	3	0.99	0.21	0.99	0.17

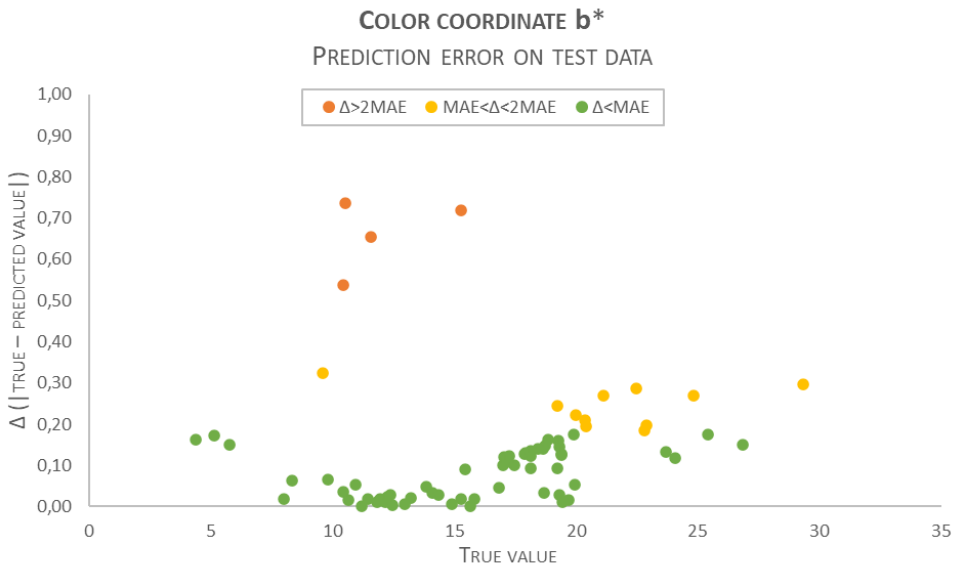

Figure 23: Prediction error in the test dataset for color coordinate b, full dataset and L*, a*, b* D65/10° as dependent variables.*

FROM COMPOSITION TO MELTING RANGE

Solidus

In the case of the solidus temperature value, the model trained with the complete dataset and having all potentially compositional elements as independent variables already exhibits good predictive capability (Table 28)

Table 28: Training parameters and prediction evaluation for solidus temperature, full dataset

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
312+ 132	19	0.83	13	0.86	13

However, observing the error distribution for the test set (Figure 24), a non-random trend is noticeable, with higher errors at the extremes of the solidus temperature value range. This may indicate a difficulty in generalizing predictions for compositions that actually have particularly high or low solidus values.

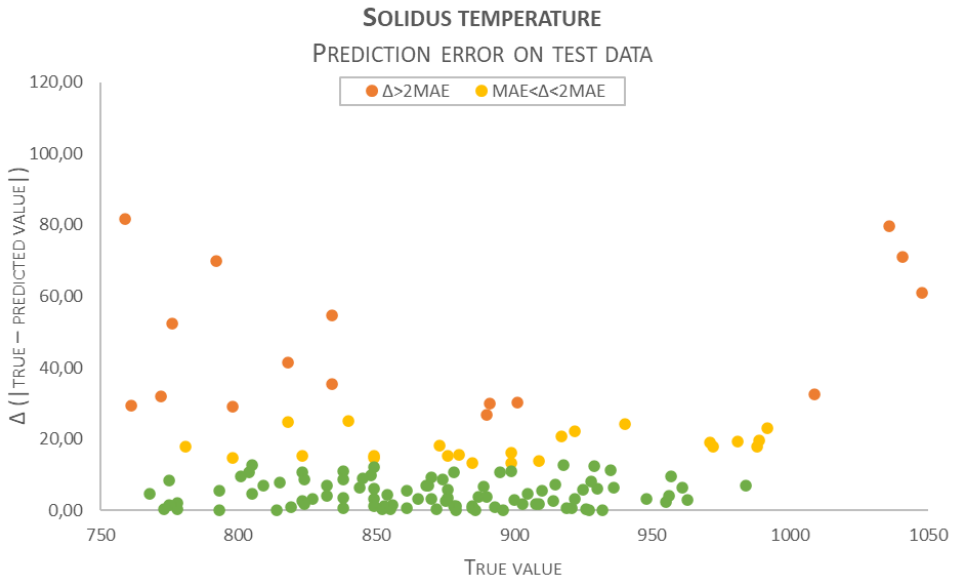

Figure 24: Prediction error in the test dataset for solidus temperature, full dataset

By reducing the number of features from 19 to 13, eliminating elements from the list of independent variables that do not contribute to the solidus value or are poorly represented, the R^2 and MAE values improve, especially in the test set (Table 29)*Table 29*

Table 29: Training parameters and prediction evaluation for solidus temperature, full dataset and reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
312+ 132	13	0.88	12	0.95	5

Furthermore, the error distribution (Figure 25) exhibits a much more random trend, indicating an improved predictive capability, especially for alloys with solidus values at the lower or upper limit of the range, which were less accurately predicted in the initial training.

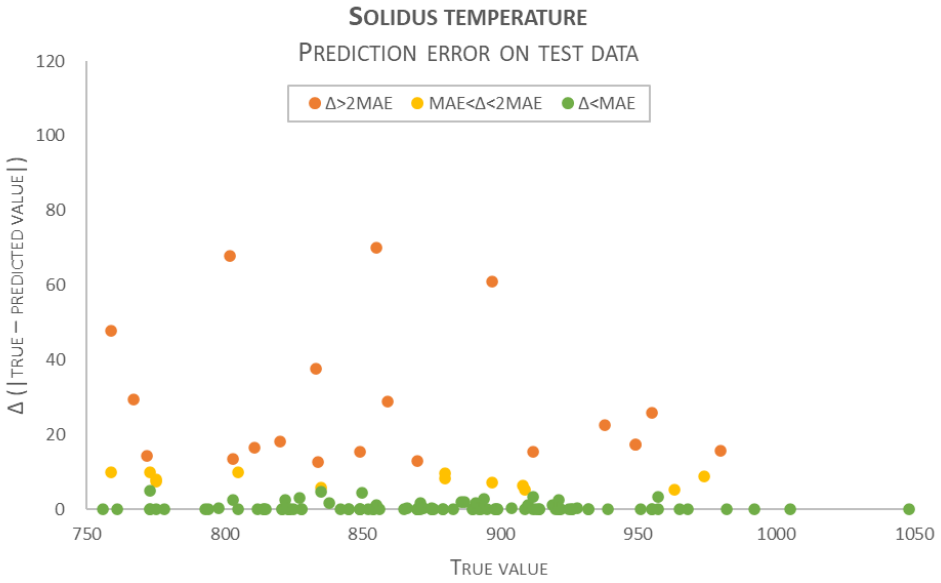

Figure 25: Prediction error in the test dataset for solidus temperature, full dataset and reduced features

Liquidus

As already observed for the solidus, in the case of the liquidus, training performed on the entire dataset, using the full range of features, shows good predictive values (Table 30, Figure 26).

Table 30: Training parameters and prediction evaluation for liquidus temperature, full dataset

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
312+ 132	19	0.91	9	0.91	11

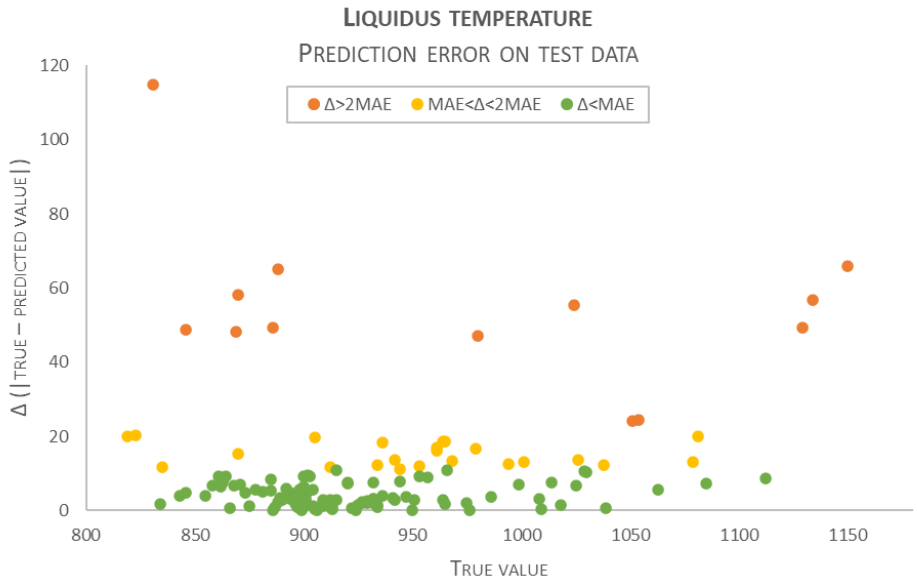

Figure 26: Prediction error in the test dataset for liquidus temperature, full dataset

Proceeding with training using reduced features results in a further improvement of evaluation parameters for the test set, while the values for the training set remain almost unchanged (Table 31).

Table 31: Training parameters and prediction evaluation for solidus temperature, full dataset and reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
312+ 132	13	0.91	10	0.97	4

In the test set, there is no longer a single point with a large prediction error (Figure 27). The alloy that had a high error in the previous training no longer exhibits an anomalous error in this case.

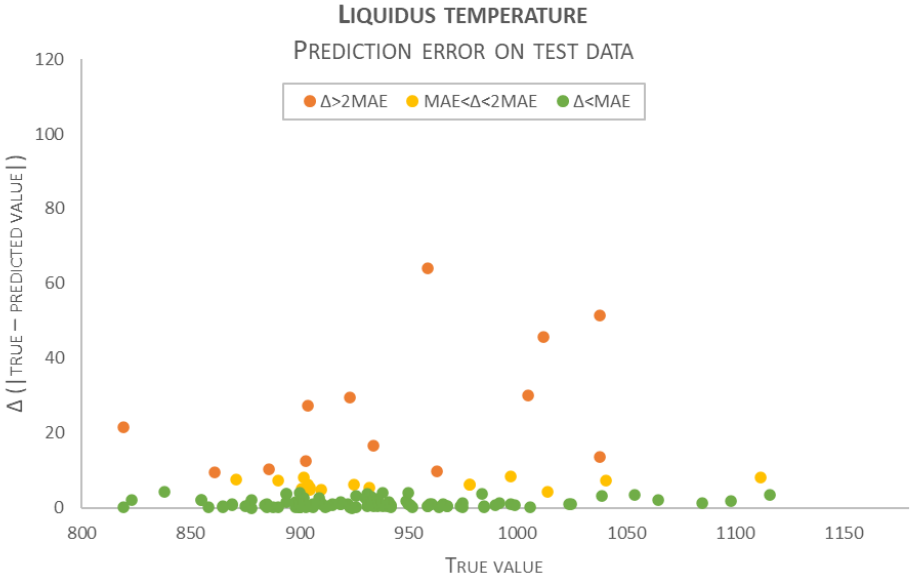

Figure 27: Prediction error in the test dataset for liquidus temperature, full dataset and reduced features

FROM COMPOSITION TO MECHANICAL PROPERTIES AFTER ANNEALING

In the case of mechanical properties derived experimentally from wire tensile tests, where the alloy is cold-worked and then annealed, it is observed that with the complete dataset and non-reduced features, the R^2 values for both the train and test sets are generally lower than those obtained for the other characteristics analyzed so far. This could be due to various reasons, such as a reduced number of data points compared to the variability of tensile behaviors of the various alloys or high experimental error that introduces uncertainty in the data used for training.

ELONGATION

Analyzing the R^2 and MAE values obtained for the prediction of maximum elongation, the described trend is evident (Table 32, **Figure 28**Figure 28): both for the train and test sets, the R^2 values are just above 0.5, indicating that slightly more than half of the variance is explained by the training model.

Table 32: Training parameters and prediction evaluation for elongation, full dataset

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	19	0.57	3	0.53	3.4

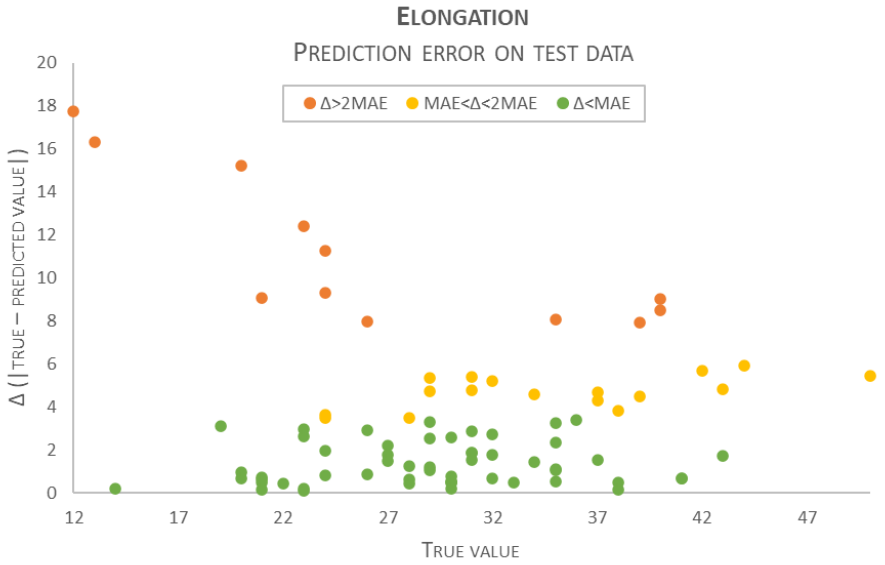

Figure 28: Prediction error in the test dataset for elongation, full dataset.

Trying to reduce the number of features (Table 33, Figure 29) actually yields an even worse result, demonstrating that in this case, the difficulty of prediction does not lie in the high number of features (i.e., possible alloying elements) but is likely due to a combination of sparse data and high experimental error.

Table 33: Training parameters and prediction evaluation for elongation, full dataset and reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	14	0.57	3.2	0.36	4

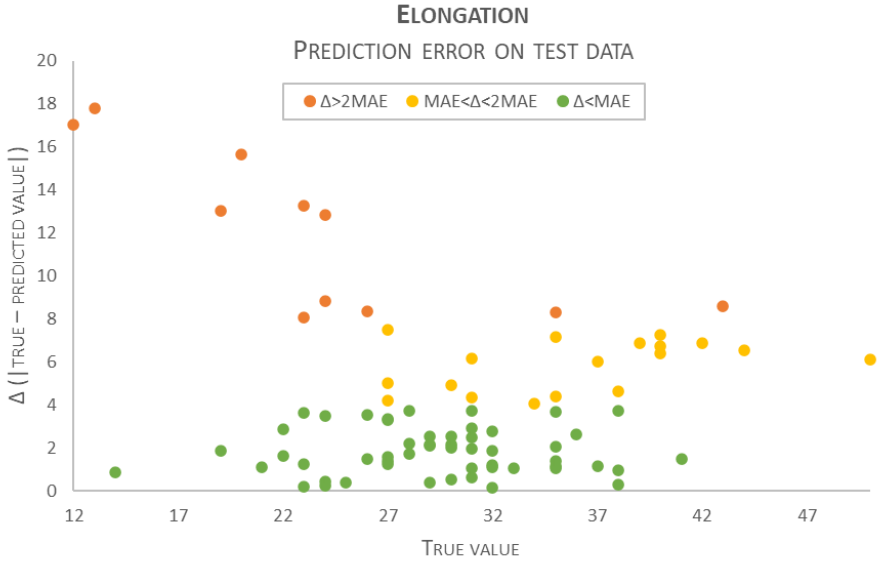

Figure 29: Prediction error in the test dataset for elongation, full dataset, reduced features.

YIELD STRENGTH

The predictive result for the physical characteristic of yield strength after annealing (Table 34, Figure 30) is better than that of maximum elongation but still remains poor compared to the other physical characteristics analyzed.

Table 34: Training parameters and prediction evaluation for yield strength, full dataset

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	19	0.59	42	0.73	32

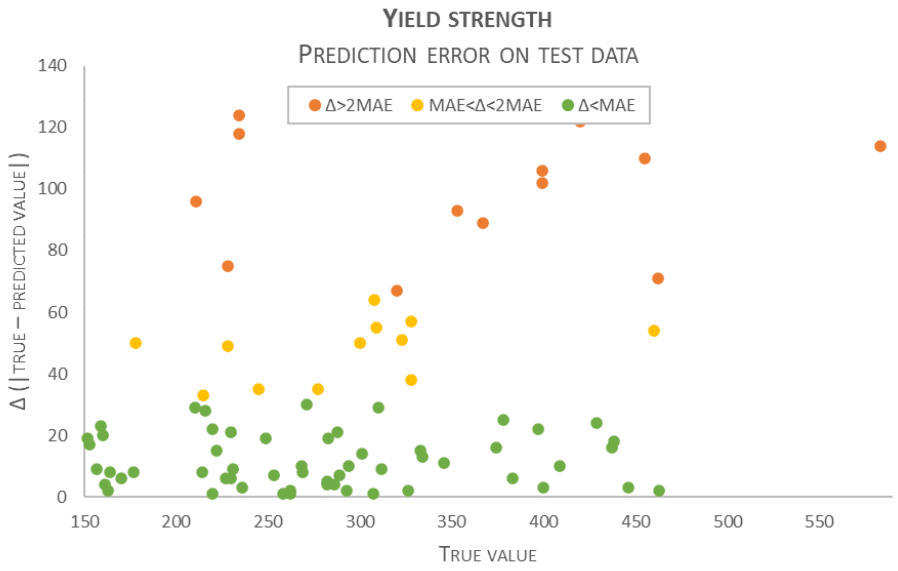

Figure 30: Prediction error in the test dataset for yield strength, full dataset.

Contrary to the previous case, however, reducing the features, identical to that performed for maximum elongation, brings a significant improvement in predictive terms (Figure 31), bringing the R^2 values (Table 35) closer to those seen for the prediction of hardness, color, and melting range.

Table 35: Training parameters and prediction evaluation for yield strength, full dataset and reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	14	0.72	29	0.88	19

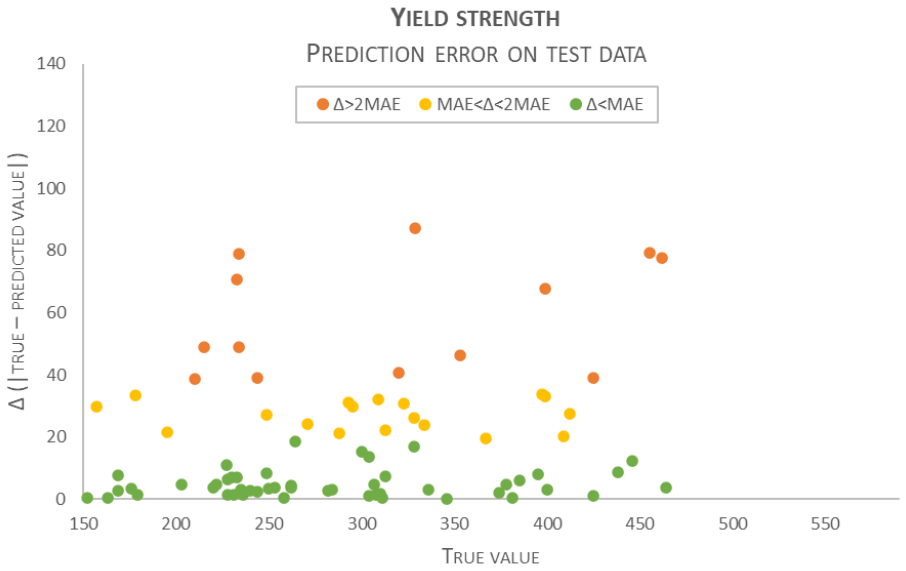

Figure 31: Prediction error in the test dataset for yield strength, full dataset, reduced features.

ULTIMATE TENSILE STRENGTH

The last mechanical property for which a prediction model was trained is the ultimate tensile strength. The results obtained are similar to those observed for yield strength: not particularly high predictive capability with all features (Table 36, Figure 32), which significantly improves by eliminating elements from the variables that do not influence mechanical behavior or are poorly represented (

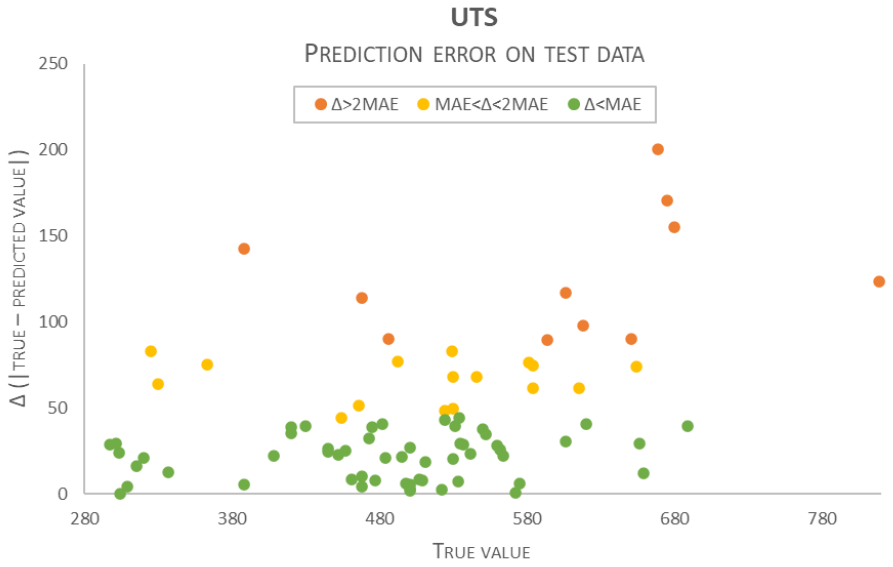

Table 37, Figure 33).

Table 36: Training parameters and prediction evaluation for UTS, full dataset

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	19	0.61	45	0.68	45

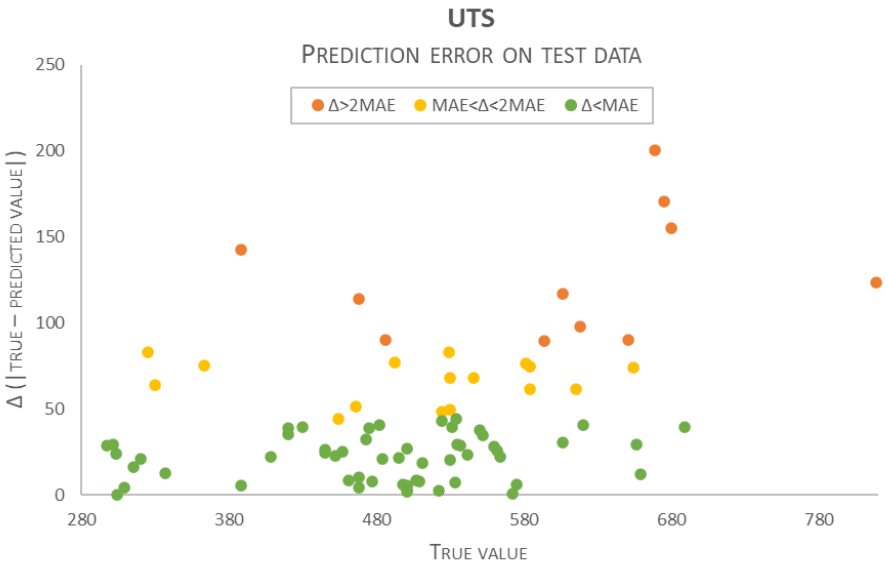

Figure 32: Prediction error in the test dataset for UTS, full dataset

Table 37: Training parameters and prediction evaluation for UTS, full dataset and reduced features

DATA N° (TRAIN + TEST)	FEATURES N°	R^2 TRAIN	MAE TRAIN	R^2 TEST	MAE TEST
192 + 83	14	0.71	29	0.83	27

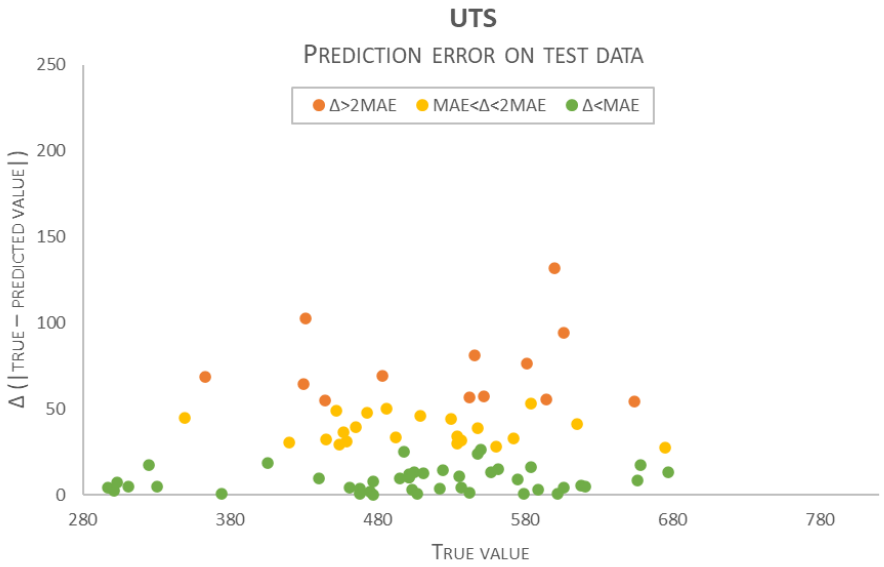

Figure 33: Prediction error in the test dataset for UTS, full dataset, reduced features.

EXPERIMENTAL TESTING

After training the models, their effectiveness in prediction was tested on the physical characteristics of a test alloy with a composition that had never been studied, and thus was not present in any of the previously used datasets. The constituent elements of the alloy are nevertheless part of the set of elements explored with previous compositions, albeit in different quantities, to stay within the range of characterized compositions.
The results obtained, both through the classical characterization process and using the trained predictive models, are reported in

Table 39, while in Table 38 are listed the trained model used to predict each property.

Table 38: trained models used for the prediction of each characteristic.

	BEST TRAINED MODEL
HARDNESS (ANNEALED)	FULL DATA
HARDNESS (HARDENED)	FULL DATA, REDUCED FEATURES
L* (D65/10°)	FULL DATA, REDUCED FEATURES
A* (D65/10°)	FULL DATA, REDUCED FEATURES
B* (D65/10°)	FULL DATA, REDUCED FEATURES
SOLIDUS	FULL DATA, REDUCED FEATURES
LIQUIDUS	FULL DATA, REDUCED FEATURES
ELONGATION (ANNEALED)	FULL DATA
YIELD STRENGTH (ANNEALED)	FULL DATA, REDUCED FEATURES
UTS (ANNEALED)	FULL DATA, REDUCED FEATURES

Table 39: comparison between experimental and predicted values for test alloy.

	EXPERIMENTAL	PREDICTED	Δ
HARDNESS (ANNEALED)	157 HV	178 HV	21 HV
HARDNESS (HARDENED)	288 HV	300 HV	12 HV
L* (D65/10°)	86,6	86,54	0,06
A* (D65/10°)	8,35	7,38	0,97
B* (D65/10°)	18,49	18,43	0,06
SOLIDUS	883 °C	882 °C	1 °C
LIQUIDUS	887 °C	892 °C	5 °C
ELONGATION (ANNEALED)	34 %	33 %	1 %
YIELD STRENGTH (ANNEALED)	333 MPa	337 MPa	4 MPa
UTS (ANNEALED)	510 MPa	507 MPa	3 MPa

Overall, the predictive capability proves to be very good for almost all explored parameters, except for coordinate a* and post-annealing hardness, which exhibit an error almost double compared to the MAE obtained in the model test.
Notwithstanding this, the prediction provides a solid preliminary indication of the alloy's characteristics in a very short time saving considerable characterization effort.
In fact, the standard characterization process for the physical properties evaluated starts with the material casting to obtain semi-finished products that are then used to produce a wire by drawing and medallions by rolling and forming. The wire is subjected to a tensile test, from which the elongation, yield strength and UTS values are derived. The medals, on the other hand, undergo various heat treatments to obtain the hardness values after annealing and after age hardening. Colour coordinates are also measured on a polished medal and finally a fragment of the material is used for thermal analysis.
The comprehensive analysis of average working times using standard characterization or simulation is reported in Table 40.

Table 40: working times for operators and instruments (minutes).

PROCESS	PHASE	OPERATOR TIME (min)	INSTRUMENT TIME (min)
STANDARD CHARACTERIZATION	ALLOY MELTING	30	30
	WIRE PREPARATION	60	60
	MEDALS PREPARATION	60	60
	COLOR TEST	20	20
	TENSILE TEST	90	90
	AGE HARDENING TESTS	120*	600*
	DTA ANALYSIS	20	180
	TOTAL	**400**	**1040**
SIMULATION	SIMULATION	20	5
	TOTAL	**20**	**5**

The time indicated for age hardening tests (600 minutes) is that required to complete the usual array of treatment conditions, e.g. temperatures of 250, 300 and 350°C for 1, 2 or 3 hours, for a total of 9 tests. At the end of these tests, not only is the maximum hardness value known, but also the temperature and time at which it was reached. At the moment, with machine learning, only the maximum hardness has a trained model and is therefore

predictable. However, it will be possible to obtain temperature and time once a specific model has been developed.
Summing up the times reported in Table 40, in the case of conventional characterization, we obtain 400 minutes of operator time, compared to 20 minutes of work using machine learning, a time 20 times lower. The ratio is even more advantageous for machine learning when considering equipment usage times. Moreover, the indicated timelines for characterization only reflect the operational times, without taking into consideration common sources of delay present in the daily laboratory activity, such as the availability of precious materials and the occupation of equipment for other projects developed in parallel.
Directly linked to the working times, the number of necessary instruments, and their energy consumption is also the environmental impact of the activities. An approximate calculation of the environmental impact of the alloy characterization, considering only the necessary processes and not the equivalent CO_2 produced by the materials extraction or refining, is presented in Table 41 (greenhouse gas emission values for electricity production in the Italian market obtained from the ISPRA report 2019 [7]). As expected, greenhouse gas emissions are in the order of grams for the simulation of the properties of a single alloy, while they reach more than 14 kg of CO_2 for a standard characterisation process.

Table 41: CO_{2eq} (Kg) produced for standard characterization and simulation.

PROCESS	PHASE	CO_{2eq} (Kg)
STANDARD CHARACTERIZATION	ALLOY MELTING	3,6
	WIRE PREPARATION	0,4
	MEDALS PREPARATION	0,1
	COLOR TEST	0,1
	TENSILE TEST	0,1
	AGE HARDENING TESTS	8
	DTA ANALYSIS	2,4
	TOTAL	**14,7**
SIMULATION	SIMULATION	0,01
	TOTAL	**0,01**

CONCLUSIONS

As a result of the tests performed, it can be stated that feature reduction generally leads to consistent enhancements in model performance metrics, such as R^2 values and mean absolute error, across diverse predictive tasks. However, it's important to consider the inherent experimental uncertainties and sensitivities of each property when interpreting the results. To properly compare prediction uncertainty with experimental uncertainty, however, we must consider the Root Mean Squared Error (RMSE) of the model, rather than the Mean Absolute Error (MAE). RMSE is indeed calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum\left(y_{true} - y_{pred}\right)^2}$$

In comparison to the formula for calculating MAE, as seen in the section on model evaluation metrics, the formula for calculating RMSE is much more similar to the formula for standard deviation (σ) used to express experimental uncertainty:

$$\sigma = \sqrt{\frac{1}{n}\sum\left(y_{measured} - y_{average}\right)^2}$$

For the models identified as the best based on R^2 and MAE, the value of RMSE for test data has also been calculated and compared to experimental standard deviation (Table 42).

Table 42: MAE and RMSE values for the test dataset of best trained model. The last column represents the standard deviation typically observed in experiments for the analyzed property

	BEST TRAINED MODEL	MAE	RMSE	σ
HARDNESS (ANNEALED)	FULL DATA	9	13	15
HARDNESS (HARDENED)	FULL DATA, REDUCED FEATURES	11	17	15
L* (D65/10°)	FULL DATA, REDUCED FEATURES	0,57	0,86	0,15
A* (D65/10°)	FULL DATA, REDUCED FEATURES	0,34	0,55	0,1
B* (D65/10°)	FULL DATA, REDUCED FEATURES	0,43	0,80	0,1
SOLIDUS	FULL DATA, REDUCED FEATURES	5	13	5
LIQUIDUS	FULL DATA, REDUCED FEATURES	4	10	5
ELONGATION (ANNEALED)	FULL DATA	3	5	3
YIELD STRENGTH (ANNEALED)	FULL DATA, REDUCED FEATURES	19	28	10
UTS (ANNEALED)	FULL DATA, REDUCED FEATURES	27	38	10

Prediction error for material hardness with reduced feature sets have resulted in errors of 13 HV for annealed alloys and 17 HV for age hardened alloys. These errors are deemed satisfactory considering the inherent experimental variability, typically around 15 HV. This improvement indicates that the selected features effectively capture the pertinent underlying factors influencing hardness variations across various alloy states.
When predicting color coordinates instead, particularly L*, a*, and b* values, the errors obtained are exceeding the experimental error. Experimental inaccuracies are in fact approximately 0.1 for a* and b* and slightly higher for L*, mainly because the lightness index is more sensitive to sample preparation techniques. The visual sensitivity of the human eye enables the perception of overall differences in terms of ΔE_{CMC} of about 1, and this value could be easily exceeded by the combination of prediction errors of the three different color coordinates. This implies that relying solely on machine learning prediction could potentially assess alloys as equal when, in reality, they exhibit visibly different colors.
For the change in color coordinates, the experimental error is zero because it is a recalculation of the values from the measured curve. For the prediction, however, there is still an error because we are approximating a calculation made from several inputs (light intensity at different wavelengths) using only 3 independent variables for training. It is worth noting that, in this case, training with maximum feature reduction (1 feature) does not correspond to the maximum prediction value, highlighting that when features are reduced too much, removing crucial information, the prediction result worsens.
As regards melting range estimation, errors on solidus (13 °C) and liquidus (10°C) are bigger than experimental errors. However, unlike with color coordinates, an error of 10°C in temperature estimation does not pose any particular practical problems and could be considered a satisfactory result. Finally, the error in predictions of ultimate tensile strength (UTS) and yield strength is slightly larger than the experimental error, which averages around 10 MPa. In the case of maximum elongation, the error of 5 on the percentage is not too far from the experimental error 3. However, it is interesting to note that in this case, feature reduction did not yield the best results.
As demonstrated by the characterization of the test alloy, the RMSE value alone provides only an indication of predictive error because, when considering each individual prediction, errors can vary widely. Prediction errors for test alloy were in fact on average much lower than RMSE values for trained models, with the exception of a* coordinate. Overall, we can affirm that the results obtained in predicting the test alloy were good, providing a solid preliminary indication of the alloy's characteristics in a very short time and saving considerable characterization effort.
To further enhance predictive performance, the primary approach would be to increase the available data for model training, particularly for compositions that are currently poorly characterized. Additionally, it's important to note that this work was carried out by non-professional machine learning operators, and conducting a professional-level analysis could certainly lead to improved results compared to the current ones.
For the future, another compelling application of trained models involves implementing a program capable of employing these models to conduct a reverse prediction process, contrasting with the one studied thus far. Rather than deriving the value of a physical property from the composition, this program would begin with a specific desired value for a physical property and predict the composition necessary to attain it. This process, although

more complex to set up, would have very interesting and immediate applications in the day-to-day work of researching new compositions. It could streamline the material design process by allowing researchers to specify desired material properties and automatically generate compositions that meet those specifications.

REFERENCE

1. Logan Ward et al., "A general-purpose machine learning framework for predicting properties of inorganic materials", *npj Computational Materials*, (August 2016): 2-7.
2. Jonathan Scmidt et al., "Recent advances and applications of machine learning in solid state materials science", *npj Computational Materials*, (August 2019): 5-83.
3. Rampi Ramprasad et al., "Machine learning in materials informatics: recent applications and prospects", *npj Computational Materials*, (December 2017): 1-13.
4. Shalev-Shwartz and Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
5. Pedro Domingos, "A Few Useful Things to Know About Machine Learning", *Communication of the ACM*, 55(10), 78-87
   https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
6. Azure machine learning documentation
   https://learn.microsoft.com/en-us/azure/machine-learning/?view=azureml-api-2
7. ISPRA, "Fattori di emissione atmosferica di gas a effetto serra nel settore elettrico nazionale e nei principali Paesi Europei", 303/2019
   https://www.isprambiente.gov.it/it/pubblicazioni/rapporti/fattori-di-emissione-atmosferica-di-gas-a-effetto-serra-nel-settore-elettrico-nazionale-e-nei-principali-paesi-europei