Supervised learning algorithms as a tool for archaeology: classification of ceramic samples described by chemical element concentrations

G. Ruschioni[1], D. Malchiodi[2,3], A. M. Zanaboni[2], L. Bonizzoni[1]*

[1] Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Italy
[2] Dipartimento di Informatica & DSRC, Università degli Studi di Milano, Italy
[3] CINI National Lab. on Artificial Intelligence and Intelligent Systems (AIIS), Italy

Corresponding author: Letizia Bonizzoni, Dipartimento di Fisica, via Celoria 16, Milano, Italy letizia.bonizzoni@mi.infn.it

Keywords

Machine learning, Supervised classification, Provenance, Ancient pottery, ED-XRF

Highlights

- Ceramics provenance investigations are of great importance in archaeological studies
- Chemical analyses coupled with statistics are useful tools for this task
- Machine Learning techniques prove to be indicated for real archaeological datasets
- Outputs provide a reliable and schematic picture of archaeological data

Abstract

Ceramic provenance studies often use minor and trace elements to gather knowledge about the presence of local furnaces and commercial trades. There are various chemical techniques that can be used to determine the elemental composition of ceramics, either non-destructively or by requiring samples. From these data, researchers can often determine provenance, and then use multivariate analyses with geological and archaeological information to classify the ceramics. In this study, we aimed to demonstrate the potential of supervised Machine Learning techniques to classify ceramic samples based on their chemical element concentrations. We applied several supervised learning algorithms to a set of 36 fragments whose archaeological classification was already known, using chemical analysis data that had been verified through previous studies. We carried out different sets of experiments, exploiting in different ways the available data, and evaluated the performance of the adopted algorithms, to propose new tools for ceramics provenance studies in archaeology. Our results show that machine learning can be a reliable and useful tool for archaeological classification based on chemical analysis data, providing a reliable and schematic picture of archaeological findings.

1. Introduction

Reliable provenance classification is based on interdisciplinary studies involving both the scientific and humanistic fields. The determination of ceramics "fingerprints" is one of the features in this complex path, and it involves several aspects, among which the identification of raw materials and the determination of manufacturing techniques (Cuomo di Caprio, 2017). These two matters are strictly linked, as raw clay was submitted to levigation and then added with tempering materials, such as chamotte or sand (Tite et al., 2003). Thus, to reach a reliable classification, a number of techniques should be synergically used, to have information about composition in both terms of chemical elements and of mineralogic phases (Sciau et al., 2015).

The examination of elemental chemical composition, with particular reference to trace elements and in association with chemometric analysis, is a relevant analytical tool used to find out different geographical provenances among the sets of archaeological pottery (Jones, 1986). Indeed, clays can have a different composition within the same quarry or, otherwise, be quite similar in different sites, hence it is in general necessary to pay particular attention to minor and trace elements (Neff, 2000). It can also be useful to juxtapose elemental and mineralogical data, the latter being strongly influenced by production techniques, especially firing temperature (Bruni et al., 2001).

Some of the most exploited techniques in the chemical characterization of ancient ceramics are: atomic emission spectroscopy (AES) (Bellanti et al., 2008), proton-induced X-ray emission (PIXE) (Robertson et al., 2002), X-ray fluorescence (XRF) (Padilla et al., 2006; Cariati et al., 2003), inductively coupled plasma (ICP) (Kennett et al., 2002), neutron activation analysis (NAA) (Descantes, 2002; Bishop, 2002), Raman and IR spectroscopy (Bruni et al, 2001), X-Ray diffraction (XRD) (Ballirano et al., 2014), Mössbauer spectroscopy (Wagner et al., 1999), and Laser Ablation Inductively Coupled Plasma Mass spectroscopy (LA-ICP-MS) (Li et al., 2006). Among these techniques, pXRF (portable X-ray Fluorescence) plays an important role, as it allows to perform non-destructive and non-invasive analyses, with short measuring times and directly in the conservation sites (Ruschioni et al., 2022). On the other hand, it does not allow the determination of the light elements matrix nor to distinguish among clay composition, inclusions and tempering materials (Frahm, 2018). For this reason, it is often used as the first approach in the Heritage material surveys, especially along with mineralogic techniques such as XRD, FT-IR or thin section studies. Hence, it is an important tile, but it gives a partial insight of materials, as many other techniques in the field of material sciences. Moreover, being pXRF a relatively superficial technique performed on small areas

of the samples, great attention must be devoted to the choice of the measuring point, as well as to the analysis of the ceramic surface, especially when they are decorated or they underwent to a burial period, as by definition the measure would not suitably represent the ceramic under study. Therefore, pXRF is usually coupled with complementary mineralogical techniques, and a subsequent deep and careful data analysis is required in order to help a correct archaeological interpretation of the obtained results. The present paper is devoted to the specific step of data handling; in other words, it focuses on the research question: "How Machine Learning techniques can help data interpretation, if they can do this?". It is worth pointing out that these techniques can be applied to all sorts of data (Anglisano et al., 2020), so that a positive answer to the previous question would also allow to treat results from different analytical techniques in one unique elaboration (Saleh et al, 2020).

2. Computer Science applied to Cultural Heritage Materials

Several methodologies in the realm of Computer Science have been recently applied to Cultural Heritage and archaeological materials, demanding cross-disciplinary cooperation and two-way communication at various levels, from data handling to museum promotion. One of the most explored fields is that related to data analysis, as statistical methods are indeed used throughout the whole archaeological research process, from the survey planning to sampling and data collection. Whenever archaeometric data are involved, the problem of data handling and analysis arise to answer specific archaeological questions; in this case, the classical approach on a wide variety of archaeological materials consists in using unsupervised methods such as Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA) and K-Means Clustering (Amadori et al., 2017; Bonizzoni et al., 2009; Bruni, 2022; Fermo et al., 2016; Galli et al., 2011). Ceramic provenance studies play an important role in gathering knowledge of local furnace presence and commercial trades: pottery sherds are the most abundant materials in archaeological excavations and archaeological queries about ceramic provenance represent a fundamental part to reconstruct the past. The examination of the elemental chemical composition in association with statistical analysis helps to find out different geographical provenances, allowing to confirm the existence of fabric groups and supporting the hypothesis of a common origin for some fragments (Bruno et al., 2000; Jones, 1986). Clays can have a different composition within the same quarry and, on the other hand, be quite similar in different sites; for this reason, it is generally necessary to pay particular attention to minor and trace elements (Neff, 2000). Indeed, multivariate statistical analysis is also generally applied to provenance ceramics studies, as a univariate study would be inadequate (Fermo et al., 2008; Liritzis et al., 2020; Papageorgiou, 2020).

For the development of an adequate model for data classification inferred (learned) from a set of available examples, Machine Learning methods can be considered. In particular, if for training examples the desired outputs are known, then supervised methods can be applied; on the other hand, if the desired outputs are unknown and we wish to find possible groupings of data based just on their similarity, unsupervised methods can be applied; a systematic survey of Machine Learning algorithms for data science can be found in (Alloghani et al., 2020). A good review of these multivariate methods and a discussion of their advantages with respect to classical methods in archaeometry can be found in (Baxter, 2006). Since that review, dating to about fifteen years ago, in which the author observed that "the explosion of interest in alternatives to the 'classical' methods of learning [….] has left the archaeometric literature largely untouched", the panorama has changed and some work has been done, in particular focusing on the classification of samples described by chemical element composition.

In (Charalambous et al., 2016), authors describe a robust methodology for choosing the best algorithm for the classification of archaeological ceramic samples coming from Cyprus; samples are described by a set of chemical compounds whose elemental concentrations were obtained through ED-XRF (Energy Dispersive X-Ray Fluorescence) analysis, the same technique used in the present paper. The data set consists of 177 measurements, a significant number for archaeological studies and, most important, similar to that considered in the present paper (112 XRF spectra from which elemental evaluation was obtained). In these experiments, K-Nearest Neighbours, Learning Vector Quantization and Decision Trees are compared. In particular, the authors highlight how the analysis of archaeological ceramic artefacts by means of classification algorithms can help to answer archaeological questions and, therefore, to identify possible typological categorizing errors or to recognize particular compositional, technological or stylistic patterns. In (Hazenfratz et al., 2017) self-organising maps are applied for the clustering of pottery shards coming from two archaeological sites in Central Amazon, and samples are described by the concentration of nine chemical elements, selected from a wider set by analytic quality control considerations and measured by INAA (Instrumental Neutron Activation Analysis). By comparison with patterns obtained through multivariate statistical methods, the authors verified the potential of Self-Organising Maps for the analysis of archaeometric data. In (Jasiewicz et al., 2021) soft clustering with Gaussian Mixture Models are combined in order to select the most important elements and classify prehistoric ceramics. Element concentrations were obtained through ED-XRF analysis and 15 elements were considered. The authors use an approach typical of supervised analysis, i.e., the selection of important variables, but applied it to unsupervised methods, minimising the disparity between the elemental classes and the position of the source material. In (Sun et al., 2020) Random Forest was the best performing algorithm compared to Support Vector Machines, AdaBoost and K-Nearest Neighbour in the multiclass classification of Chinese ancient ceramics; the classification model is enriched with Mahalanobis distance in order to determine how far the sample is from the centre of the predicted class, and the most relevant chemical elements for sample description are selected looking at their influence on classification accuracy. The resulting classification model was also applied in practical archaeological problems. As in (Charalambous et al., 2016) and (Jasiewicz et al., 2021), the elemental dataset used for the studies was obtained through ED-XRF analysis.

The aim of the present work is to create a model able to distinguish between fragments of Etruscan pottery classified as local production from other fragments having a different provenance. Since the performance of a learning algorithm strongly depends on the structure of data, we compared the results of several supervised learning algorithms for classification: in addition to the methods discussed in (Baxter, 2006), we also tested naive Bayes methods and Random Forests. The first methods are based on the assumption that variables describing objects are independent and exploit the Bayes' theorem for building the classification decision rule (Maritz and Lwin, 2018); on the other hand, Random Forests are an ensemble of Decision Trees and use an aggregation rule (e.g., majority vote) for outputting their prediction (Breiman, 2001).

3. Materials and methods

With the aim of showing the potential of supervised methods on real archaeological data, we have considered a set of 36 fragments, whose known archaeological classification is summarised in Table 1. This classification, which has been made on archaeological bases, has been verified through chemical methods (Bonizzoni et al., 2010; Bruni et al., 2001; Fermo et al., 2004) coupled with statistical elaborations. Among the archaeometric analysis performed, pXRF had been also considered; in this work, we start again from the same XRF spectra and elemental concentrations, originating the previous classification, to test the supervised learning algorithms proposed in the present paper. For each fragment, several measuring points were considered, as detailed later in the text and reported in Table 1; the results of all measurements (112 on the 36 fragments) were used to prepare the dataset for statistical elaboration.

We considered 27 fragments of Etruscan *depurata* pottery, with most of them belonging to the *vernice nera arcaica* (black varnish decoration) class, while the remaining ones belong to the *etrusco geometrica* (geometrical decorations), *etrusco corinzio*, and *bucchero* pottery class. They are all from the archaeological excavation at Pian della Civita in Tarquinia (Italy), classified as local production and dating from the VIII to the IV century B.C. Six additional fragments of black varnish fine pottery are from the Greek colony of Velia, dating the same period. Three further samples of non-local origin have been included, even if no hypothesis on their provenance were previously made. It is worth noting that somehow a vague classification is typical of real archaeological contexts, and thus this set of data, even though not ideal from a purely statistical point of view, is a representative and challenging case study for the aim of the present research.

The element concentrations employed for calculation presented in this work have been obtained through non-destructive quantitative X-ray fluorescence (XRF) analysis, exploiting a portable spectrometer (Bonizzoni et al., 2010); this technique has proved to be useful as a first check for the presence of the same raw materials when coupled with multivariate statistical treatment of data (Romano et al., 2006; Padilla et al., 2006; Idjouadiene et al., 2019), making analyses possible for a wide range of materials even when sampling is forbidden (Fermo et al., 2016; Galli et al., 2011; Veneranda et al., 2022). XRF measurements were performed on selected areas on untreated ceramics with a portable spectrometer (Assing Lithos 3000) equipped with a low power X-ray tube with Mo anode and a Peltier cooled Si-PIN detector; the working conditions were 25 kV and 0.3 mA with a 500 s acquisition time. From qualitative analyses of the spectra, eleven elements were detected; among these, Cu and Zr were under the minimum detection limit for most of the samples and were not considered for quantification. We thus considered the remaining nine elements, namely K, Ca, Ti, Cr, Mn, Fe, Zn, Rb and Sr, for quantitative analysis. For elements showing a concentration under Minimum Detection Limit (MDL) only in a few samples, we substituted the missing data in the overall dataset table with a random concentration value between 0 and the detection limit itself. The MDL values for trace elements were estimated considering background fluctuation and instrument sensitivity for each single spectrum; the quantitative analyses were performed using a computational method (Lithos 3000 software) based on the fundamental parameters and considering, in addition to the characteristic X-ray lines of the elements, also the intensity ratio of the scattered peaks (Bonizzoni et al., 2010), to get information on effective Z of low elements' matrix or at least on its behaviour regarding X-ray absorption. Due to the intrinsically inhomogeneous nature of ceramics, even if we are dealing with fine pottery, more than one measure was considered for each fragment (from 2 to 7, depending on the dimension and conservation state of the fragment, as reported in Table 1); the total number of measures considered was 112. Non-decorated areas and fresh fractures were preferred to minimise contamination from burial or material unrelated to the ceramic bulk. Compton normalisation has been applied to verify possible geometry problems and discard spectra before performing quantitative analysis. A reference sample with composition similar to the unknown ones was also measured to get elemental sensitivity and geometrical efficiency. For our samples, the sum of weight concentrations of detected medium-heavy elements is between 5% and 25%, depending on the samples, but it has a lower variation (a few percent) among different spots on the same sample. Considering the average concentration values for each fragment, statistical errors on calculated concentrations are about 10%, due to the local inhomogeneity of ceramic sherds. Prior to statistical elaborations, the weight concentrations obtained for all detected elements from a given spectrum have been normalised to 100 and can no longer be regarded as weight concentration. This simplifies sample comparisons, and it eliminates the differences among samples due to the varied silicate presence or firing temperatures, which could induce a different weight loss also in fragments with similar raw materials. This procedure is particularly advisable whenever samples contain indefinite amounts of extraneous material (Aruga, 1998), as the case of archaeological ceramics were extraneous substances such as crushed shell or crushed stone, called temper, could have been added to the original raw material in order to improve the properties of the manufactured products. Indeed, the applied normalisation decreases the number of variables by one: possible implications of this aspect are discussed in the following sections, when dealing with dimensionality reduction of data. It is worth noting that the number of measure points on each

fragment is not the same, as the dimension of fragments was quite different; even if this feature could lead to a slight skew of results, once again this is a typical situation when dealing with a real set of archaeological data.

Table 1: samples considered for testing supervised methods; for each sample, the number of XRF spectra/measured points is reported. Further archaeological details can be found in the paper quoted in the text.

Fragments			Number of measurement points for each fragment	Archaeological classification
180_96			6	Tarquinia (local)
170_2	3_73	48_23	4	
186_2	13_4	40_8		
193_43	3_607	274_7		
c203-1				
88_74	3_610	121_2	3	
59_35	199_33	80_25	2	
59_159	227_35	227_46		
28_66	c3-738	3_204		
28_128	c30-110	A10_25		
c281-60				
72			3	Non local
c258-12	A10-3bis		2	
V1			7	Velia
V2	V5	V6	4	
V4			3	
V3			2	

4. Statistical analysis

The available data (112 data points) were obtained from the 36 fragments listed in Table 1; 27 fragments (75%) were labelled as local production and 9 (25%) were labelled as non-local production. Taking into account the repeated measures for each fragment, we ended up with 112 data points, 81 (72.3%) labelled as local production and 31 (27.7%) labelled as non-local production. Table 2 shows the distribution of the number of repeated measures in the dataset; in particular, for all fragments at least two measures were taken, and the number of measures for a given fragment was at most 7.

Each data point is described by the relative concentration of the already mentioned nine elements K, Ca, Ti, Cr, Mn, Fe, Zn, Rb and Sr. Normality of chemical elements was tested by the Shapiro-Wilk test; some elements showed a normal distribution, in particular Ca, Ti, Cr and Mn in the local production group and K, Ca, Ti, Cr, Zn and Rb in the other group.

Table 2: frequency distribution of measures on same samples.

Number of repeated measures	Number of samples
1	0
2	16
3	5
4	13
5	0
6	1
7	1

Table 3 illustrates the results of a preliminary statistical analysis of the gathered dataset. In particular, we computed central position and dispersion values, comparing the two groups according to the Student's t test (for normal variables) or the Mann-Whitney U test (for non-normal variables); all the elements had a statistically different behaviour in the two groups ($p<0.001$ for all the significant differences), except Sr. In addition, Figure 1 shows the histograms of chemical elements in the two groups: samples coming from Tarquinia have a higher Ca concentration than samples of non-local production, and this reflects the well-known characteristic of Tarquinia raw material, rich in illitic-kaolinitic clays (all containing Ca), mostly if compared to surroundings Etruscan sites (Fermo et al., 2004). Moreover, a narrow distribution is present for Tarquinian samples, while the non-local ones are more spread, reflecting the possible different origin of some of them.

Table 3: Element concentration: central and dispersion values in the two groups, with normality and significant differences.

* normal in both groups		MEAN or MEDIAN		STDEV or IQR		T-test or Mann-Whitney U test
		LOCAL	NOT LOCAL	LOCAL	NOT LOCAL	p-value <0-001
*	**Ca**	52.13	22.69	8.33	9.03	*
	Fe	29.08	43.73	7.42	7.89	*
	K	15.7	29.28	4.72	5.59	*
*	**Ti**	2.25	3.65	0.68	0.96	*
	Mn	0.71	1.18	0.23	0.69	*
*	**Cr**	0.29	0.53	0.11	0.20	*
	Sr	0.12	0.11	0.07	0.05	
	Zn	0.07	0.12	0.03	0.07	*
	Rb	0.05	0.10	0.02	0.02	*

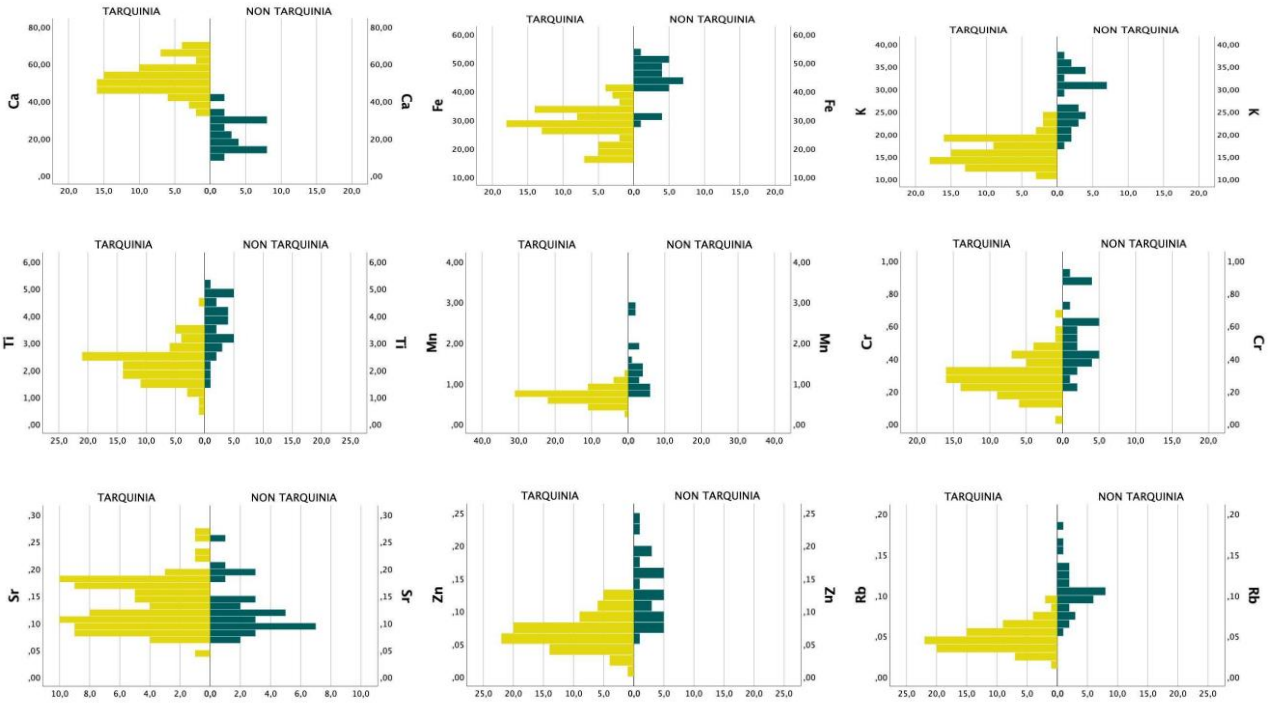


Figure 1: Histograms of element concentrations in the two classes of provenance (local and non-local).

5. Experiments

For the classification task we compared ten different supervised machine learning algorithms, namely: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Neural Networks (in particular Multi-Layer Perceptrons, MLP), Support Vector Machines (SVM) in their linear and non-linear versions (the latter based on polynomial and Gaussian kernels), binary Decision Trees (DT), Random Forests (RF), Naive Bayes (NB) and K-Nearest Neighbors (KNN)[1]. In almost all cases, we directly used the scikit-learn python library (Pedregosa et al., 2011) for running these learning algorithms: the only exception was done for MLPs, whose scikit-learn implementation was tweaked in order to deal with a single output neuron activated using a logistic function.

As we were dealing with a dataset of limited size, in order to perform model selection and jointly assessing model performance, we used the nested *k*-fold cross-validation resampling technique, and stratified training and test sets according to the two classes of local and non-local production. In simple *k*-fold cross-validation the original data set is divided into *k* folds of the same size, where one fold is used for validation and the remaining *k*-1 folds are used for training the model (in other words, data are split in a training and a test set approximately containing (*k*-1)/*k* and 1/*k* of the original dataset, respectively); the process is repeated *k* times, choosing at each time a different fold for validation, thus considering all the possible non-overlapping splits in train and validation of the original dataset. At each iteration a model is learned, and its generalisation capability is evaluated on the corresponding validation set. The average of these evaluations is used as an estimate of how any of the *k* learned models will perform on data not used during the training phase. If the learning algorithm is also characterised by a set of parameters (called hyper-parameters) that have to be tuned before model inference, as happening with almost all of the previously mentioned algorithms, nesting two *k*-fold cross-validation processes is a good method for performing model selection and evaluation: in the outer loop the entire data set is divided into several training and test sets (according to the *k*-fold technique), in the inner loop the best hyper-parameters configuration (through search, for instance, in a grid of possible values) is found by a second *k*-fold cross validation on the training set; the final model is then re-trained using the complete training set and the best values for hyper-parameters, and its generalisation capability is evaluated on the outer test set. The overall process is repeated as many times as many folds we have in the outer cross-validation. The metric we used for model selection was accuracy, namely the fraction of examples correctly classified. We used 4 external folds and 3 internal ones. Generalisation ability was measured using accuracy, sensitivity, specificity, and F1 score (see later on for their formal definition).

Despite the fact that the original data were already normalised, we tested a few dimensionality reduction techniques and a few scaling methods. Namely, concerning dimensionality reduction we exploited Principal Component Analysis (PCA) and Singular Value Decomposition (SVD), considering all the possible number of extracted components for both techniques. Scaling involved standardisation, normalisation, re-scaling to the [0, 1] interval, and robust scaling (via quantile extraction). We did not consider other scaling techniques, notably those based on logarithmic transformations, traditionally applied when dealing with statistical analysis on whole spectra, to avoid noise amplification. Scaling and dimensionality reduction are part of a pre-processing phase, but, since the choice of a particular method could perform better when coupled with a given learning algorithm and a particular choice of its hyper-parameters, we decided to incorporate them in the hyper-parameters grid search step, giving the possibility of ignoring either or both steps (thus, in the latter case, we directly process raw data). When performing dimensionality reduction, we have extracted seven variables as maximum size; in some cases, less than five dimensions were used for calculations. Please note that the two experiments for which all the variables were considered without dimension reduction showed the best performances.

The hyper-parameters we considered for model selection and the corresponding grid values are listed in Table 4. In particular, for MLPs we conjectured that at most two hidden layers were sufficient for separating the two groups; for LDA we tried different solvers; for KNNs we decided to consider at most eight neighbours, given the small number of negative examples in the dataset; for SVMs we considered the linear, polynomial and Gaussian kernels, and values for the regularisation (inverse of penalty) C parameter ranging in the logarithmic space from 10^{-4} to 10^{3}; for binary DTs we tried entropy and the heterogeneity Gini index for node splitting and allowed a maximum number of features for node conditions ranging from the total number of descriptors (9) to its square root (3); in RFs we thought that, given the binary nature of the classification problem, an odd number of estimators was preferable, and we allowed a minimum of 3 and a maximum of 9 estimators; we did not perform any model selection for NB classification, using the default implementation in scikit-learn. The software implementing our experiments and data matrix are available for replicability/reproducibility purposes at https://github.com/dariomalchiodi/JAS-Tarquinia-classification.

We performed three different types of experiments, and we will refer to them as type 1, type 2, and type 3 experiments henceforth. In type 1 experiments we used the complete dataset as if all data were independent; this means that stratification and subdivision in folds were run over all the available data points, so that all the different measures taken from each fragment contributed individually to the learning process. However, since in this case data points are "not completely" independent, we assumed that this kind of strategy might lead to overfitting. In this simple way of exploiting data, indeed, only some among the measures of a given fragment might fall in a fold used for training; in this case, the remaining measures of the same fragment would belong in the fold devoted to model selection or assessment. Therefore, train and test sets used in the cross-validation process may be "not completely" disjoint. In order to overcome this problem, we designed type 2 experiments, where stratification and subdivision in folds were done on fragments rather than on measures; in this setting, each fragment was considered just once and the fold containing the fragment contained all its available measures. The drawback here is that now folds may have different sizes, since the number of measures is not constant across fragments (see Table 2). In order to consider folds having the same size, we conceived type 3 experiments, where stratification and subdivision in folds were done on fragments in the same way as in type 2 experiments, now considering only two measures for each fragment, sampled from the available ones.

Table 4: Model, hyperparameters and grid values. For the sake of brevity, the column Hyperparameters shows the names used by the scikit-learn library.

Model	Hyperparameters	Grid values
LDA	solver	'svd', 'lqsr'
SVM	C	V = set of ten values evenly spaced between 1E-4 and 1E3 in logarithmic space
	kernel	linear, polynomial (with degree p ranging in {2, 3, 5, 9}), gaussian (with parameter γ ranging in V, also allowing the predefined 'auto' and 'scale' settings)
DT	criterion	Gini index, entropy
	max_features	square root of total number of features, no maximum number of features
	max_depth	2, …, 9, ∞
	min_samples_split	2, …, 5
	min_samples_leaf	2, …, 5
	ccp_alpha	0, 0.5, 1, 1.5
RF	same hyperparameters of DT +	
	n_estimators	3, 5, 7, 9
KNN	n_neighbors	1, …, 7
	p	2, 3
MLP	hidden_layer_sizes	one hidden layer with two neurons, one hidden layer with three neurons, two hidden layers with two neurons each
	activation	logistic, ReLU
	alpha	1E-4, 1E-3
	learning_rate	'constant', 'adaptive'
	learning_rate_init	1E-4, 1E-3, 1E-2
	shuffle	True, False
	momentum	0.8, 0.9
LR	penalty	L1 and L2 regularization
	C	V

6. Results and discussion

For each considered learning algorithm, the best performing model according to the accuracy score (i.e., the one which corresponds to the best performing hyper-parameters) found in the internal cross validation was tested on the test sets of the external cross validation. In correspondence of each data point of the test set, the model output was 1 if its predicted class was Tarquinia, and it was 0 otherwise. Since the external folds were four, we ended up with four best performing models for each learning algorithm, and we used those models for evaluating the algorithm performance. Focusing on the model maximising accuracy among such best models, Tables 5, 6 and 7 (each devoted to one of the performed experiments) display the optimal values of the considered algorithms' hyper-parameters, together with the optimal data transformation and dimensionality reduction technique. In other words, these tables report, for each experiment and for each considered model, the values of the hyper-parameters that can be used in order to train the best performing model without re-executing the model selection phase, in order to rapidly reproduce our results.

We considered different performance measures. In Tables 8, 9 and 10 (also in this case, each table describes an experiment) such measures are evaluated considering the four best performing models through mean and standard deviation (in brackets). These evaluations concern data not used during the training phase, thus they suitably summarise the generalisation capability of the induced models when they will be queried with new data. In particular, let us introduce some intermediate concepts which we will use in order to define the considered measures.

We denote as positive the samples belonging to the local production (Tarquinia) class, and as negative the remaining samples. Let us define: i) P and N, respectively, as the total number of positive and negative samples; ii) TP (true positives) as the number of positive samples correctly classified; iii) TN (true negatives) as the number of negative samples correctly classified; iv) FN (false negatives) as the number of positive samples erroneously classified. Now:

- accuracy is the ratio of correct predictions over the total number of samples, that is (TP + TN) / (P + N),
- sensitivity (also known as recall) is the analogous of accuracy when only considering positive samples, that is TP / P,
- specificity is the analogous of accuracy when only considering negative samples, that is TN / N,
- F1 score is the harmonic mean between sensitivity and precision, the latter intended as the ratio of correct positive classifications over the total number of positive predictions, that is TP / (TP + FN).

Our dataset was not balanced: in particular, the positive class was over-represented. In such cases, accuracy might not be a reliable performance estimator. This is why we also considered the remaining measures: sensitivity and specificity are specialised on a particular class, whereas F1 score synthesises a unique numerical value for both classes, with a sudden drop when performance on either one of them decreases. Summing up, Tables 8, 9 and 10 summarise for each experiment and for each model the values of the four performance-metrics listed above: more precisely, they report the mean and standard deviation on the test set for each fold of the cross-validation process. For all metrics, the higher the first value, the more performing is the corresponding model on the average. Similarly, the lower the standard deviation, the higher is the uniformity of the performances on the cross-validation folds.

The accuracy for all the considered learning algorithms in any of the three experimental settings was not less than 0.85, and the F1 score was not less than 0.88. In many cases they were higher than 0.95. All the considered algorithms performed well on the positive class (sensitivity), while some performed poorly on the negative class (specificity). We underline that in such cases the standard deviation has the same magnitude of the mean, thus the models are not highly predictive on the negative class.

Puzzlingly, type 1 experiments reached the best performances, although they were the weakest from a methodological point of view, as discussed in the previous section. However, they produced the best models because, in the actual context, different measures of the same fragments can be considered as independent observations. In fact, ceramics can present inclusions locally changing chemical composition on small areas, thus it is reasonable to consider each measure on the same object as independent. Type 1 experiments thus allow to indirectly verify the sufficient homogeneity of the fragment and thus its suitability to the classification by punctual chemical analyses, such as portable ED-XRF. Type 3 experiments performed in general worse than the other two types of experiments, and this could also be due to the fact that some data points were lost if not selected in the sampling phase, reducing the training set size.

As a general trend, non-linear models perform better than linear ones in almost all settings, although with notable exceptions: on the one hand, the score of linear SVMs is in the upper part of the ranking; on the other one, Neural Networks are always among the less performing models. We hypothesise that this is due to the high complexity of such models, which easily leads to overfitting. Indeed, very simple models such as NB and KNNs consistently ranked as best across all experiments.

Table 5. Characterisation of the best performing models found for each considered learning algorithm in experiments with stratification performed on all measures.

Model	Best parameters
NB	MinMaxScaler + PCA(n=3)
KNN	MinMaxScaler + TruncatedSVD(n=2), n_neighbors: 1, metric: minkowski, p: 2
SVM-lin	StandardScaler + PCA(n=7), C: 0.13
SVM-rbf	StandardScaler + PCA(n=4), C: 4.64, gamma: 0.0036
SVM-poly	StandardScaler + PCA(n=2), C: 166.81, degree: 5
RF	No scaling, PCA(n=2), criterion: gini, max_features: sqrt, max_depth: 9, min_samples_split: 5, min_samples_leaf: 3, ccp_alpha: 0, n_estimators: 3
LDA	No scaling, PCA(n=5), solver: svd
DT	No scaling, TruncatedSVD(n=2), criterion: gini, max_features: None, max_depth: None, min_samples_split: 2, min_samples_leaf: 2, ccp_alpha: 0
LR	StandardScaler + PCA(n=6), penalty: l2, C: 27.83, solver: liblinear, max_iter: 5000
MLP	No scaling, PCA(n=2), hidden_layer_sizes: [2], activation: logistic, alpha: 0.0001, learning_rate: constant, learning_rate_init: 0.001, shuffle: True, momentum: 0.9

Table 6. Characterization of the best performing models found for each considered learning algorithm in experiments with stratification performed on all fragments, considering all available measurements.

Model	Best parameters
NB	No scaling, no dimensionality reduction
KNN	MinMaxScaler + PCA(n=2), n_neighbors: 1, metric: minkowski, p: 2
SVM-lin	StandardScaler + PCA(n=2), C: 0.129
SVM-poly	StandardScaler + PCA(n=2), C: 1000, degree: 3
SVM-rbf	StandardScaler + PCA(n=2), C: 166.81, gamma: 0.00059
RF	No scaling, PCA(n=2), criterion: gini, max_features: sqrt, max_depth: None, min_samples_split: 5, min_samples_leaf: 4, ccp_alpha: 0, n_estimators: 3
LDA	No scaling, PCA(n=7), solver: svd
DT	No scaling, TruncatedSVD(n=2), criterion: gini, max_features: None, max_depth: None, min_samples_split: 3, min_samples_leaf: 3, ccp_alpha: 0
LR	No scaling, PCA(n=2), penalty: l1, C: 0.129, solver: liblinear, max_iter: 5000
MLP	No scaling, PCA(n=2), hidden_layer_sizes: [2], activation: logistic, alpha: 0.0001, learning_rate: constant, learning_rate_init: 0.001, shuffle: False, momentum: 0.9

Table 7. Characterization of the best performing models found for each considered learning algorithm in experiments with stratification performed on all fragments, considering each time two sampled measurements.

Model	Best parameters
NB	No scaling, no dimensionality reduction
KNN	MinMaxScaler + PCA(n=3), n_neighbors: 5, metric: minkowski, p: 2
SVM-lin	StandardScaler + PCA(n=2), C: 4.641
SVM-poly	StandardScaler + PCA(n=3), C: 4.6415, degree: 3
SVM-rbf	StandardScaler + PCA(n=2), C: 0.774, gamma: auto
RF	No scaling, PCA(n=2), criterion: entropy, max_features: sqrt, max_depth: 6, min_samples_split: 2, min_samples_leaf: 3, ccp_alpha: 0, n_estimators: 7

LDA	RobustScaler + PCA(n=2), solver: svd
DT	No scaling, TruncatedSVD(n=2), criterion: gini, max_features: None, max_depth: None, min_samples_split: 2, min_samples_leaf: 3, ccp_alpha: 0
LR	No scaling, PCA(n=2), penalty: l2, C: 0.00359, solver: liblinear, max_iter: 5000
MLP	No scaling, PCA(n=2), hidden_layer_sizes: [2], activation: logistic, alpha: 0.0001, learning_rate: adaptive, learning_rate_init: 0.001, shuffle: True, momentum: 0.9

Table 8. Performance of the best performing models found for each considered learning algorithm in type 1 experiments (stratification performed on all measures). Values outside and within brackets represent mean and standard deviation, respectively.

Model	Accuracy	Sensitivity	Specificity	F1
NB	0.98 (0.03)	1.00 (0.00)	0.93 (0.12)	0.99 (0.02)
KNN	0.97 (0.03)	1.00 (0.00)	0.90 (0.12)	0.98 (0.02)
SVM-lin	0.96 (0.03)	1.00 (0.00)	0.87 (0.10)	0.98 (0.02)
SVM-poly	0.96 (0.03)	1.00 (0.00)	0.87 (0.10)	0.98 (0.02)
SVM-rbf	0.96 (0.03)	1.00 (0.00)	0.87 (0.10)	0.98 (0.02)
RF	0.95 (0.06)	0.99 (0.02)	0.83 (0.17)	0.96 (0.04)
LDA	0.94 (0.05)	1.00 (0.00)	0.77 (0.19)	0.96 (0.03)
DT	0.91 (0.04)	0.96 (0.04)	0.77 (0.19)	0.94 (0.03)
LR	0.91 (0.05)	0.91 (0.10)	0.90 (0.12)	0.93 (0.05)
MLP	0.85 (0.07)	0.81 (0.13)	0.93 (0.12)	0.88 (0.06)

Table 9. Performance of the best performing models found for each considered learning algorithm in type 2 experiments (stratification performed on all fragments, considering all available measurements). Same notations as in Table 8.

Model	Accuracy	Sensitivity	Specificity	F1
NB	0.98 (0.04)	1.00 (0.00)	0.88 (0.22)	0.99 (0.02)
KNN	0.96 (0.03)	0.99 (0.02)	0.86 (0.21)	0.98 (0.02)
SVM-lin	0.95 (0.05)	1.00 (0.00)	0.78 (0.22)	0.97 (0.03)
SVM-poly	0.93 (0.06)	1.00 (0.00)	0.70 (0.27)	0.96 (0.03)
SVM-rbf	0.96 (0.04)	1.00 (0.00)	0.81 (0.21)	0.98 (0.02)
RF	0.90 (0.06)	0.92 (0.08)	0.82 (0.21)	0.93 (0.05)
LDA	0.90 (0.09)	0.99 (0.02)	0.61 (0.40)	0.94 (0.05)
DT	0.85 (0.09)	0.87 (0.08)	0.73 (0.28)	0.90 (0.06)
LR	0.89 (0.07)	0.93 (0.11)	0.69 (0.32)	0.92 (0.06)
MLP	0.89 (0.06)	0.94 (0.06)	0.63 (0.41)	0.93 (0.03)

Table 10. Performance of the best performing models found for each considered learning algorithm in type 3 experiments (stratification performed on all fragments, considering each time two sampled measurements). Same notations as in Table 8.

Model	Accuracy	Sensitivity	Specificity	F1
NB	0.96 (0.07)	0.96 (0.06)	0.94 (0.11)	0.97 (0.05)
KNN	0.97 (0.05)	1.00 (0.00)	0.88 (0.22)	0.98 (0.03)
SVM-lin	0.99 (0.02)	1.00 (0.00)	0.96 (0.07)	0.99 (0.02)
SVM-poly	0.93 (0.05)	0.98 (0.04)	0.75 (0.25)	0.96 (0.03)
SVM-rbf	0.94 (0.04)	1.00 (0.00)	0.77 (0.18)	0.96 (0.02)
RF	0.93 (0.07)	0.98 (0.03)	0.75 (0.25)	0.96 (0.04)
LDA	0.92 (0.05)	0.98 (0.04)	0.71 (0.22)	0.95 (0.03)
DT	0.92 (0.06)	0.93 (0.09)	0.88 (0.22)	0.94 (0.04)
LR	0.85 (0.05)	0.83 (0.11)	0.88 (0.22)	0.89 (0.04)
MLP	0.88 (0.07)	0.87 (0.10)	0.88 (0.22)	0.91 (0.05)

As far as dimensionality reduction is concerned, in most cases the best choice was the use of only two components (by PCA or truncated SVD). This was unexpected since, as discussed in Section 3, the chemical elements behave quite differently in the two classes and the first two PCA components explain only 75% of total variance[1]. It is worth noting that a relatively low variance is often found for ceramics classifications through the joint use of XRF and PCA (Frahm, 2018; Freitas et al., 2018; Liritzis et al., 2020). However, the first two PCA components seem to be sufficient for a good separation of the two classes, as shown in Figure 2. We highlight that the group separation reported is not evident as the obtained classes do not form compact and well separated groups. Nonetheless, the learning algorithm is able to find an efficient discriminant function for grouping the measurements, as clear from the figures reported in Appendix A: supplementary materials. It must be taken into account that this is not the output of the data elaboration, but an intermediate step before performing the actual classification; the final output will be a direct answer for the pertaining group. Table 11 shows the coefficients of the PCA transformation (computed on the overall dataset) for the first and the second components, ordered by decreasing absolute value. It can be noted that Ca is the element contributing with the highest absolute weight in the first component, while in the second component it is Sr.

Table 11. Coefficients for the first and the second PCA components, ordered by decreasing absolute value

First principal component		**Second principal component**	
Ca	-0,965	Sr	0,934
Fe	0,898	Rb	0,381
Rb	0,827	Zn	0,322
Ti	0,812	Cr	-0,165
K	0,807	Fe	-0,164
Mn	0,77	Ca	0,119
Zn	0,743	Ti	0,113

1 One of the reviewers pointed out that in the dimensionality reduction phase a robust estimation of location and spread could be more appropriate than standard PCA, and suggested the use of the R rrcov package (Todorov 2022). We found that, in our specific dataset, the outliers detected by the more robust methods didn't represent a difficulty for the classification task, as indeed Figure 2 shows.

Cr	0,732	K	-0,07
Sr	-0,212	Mn	-0,012

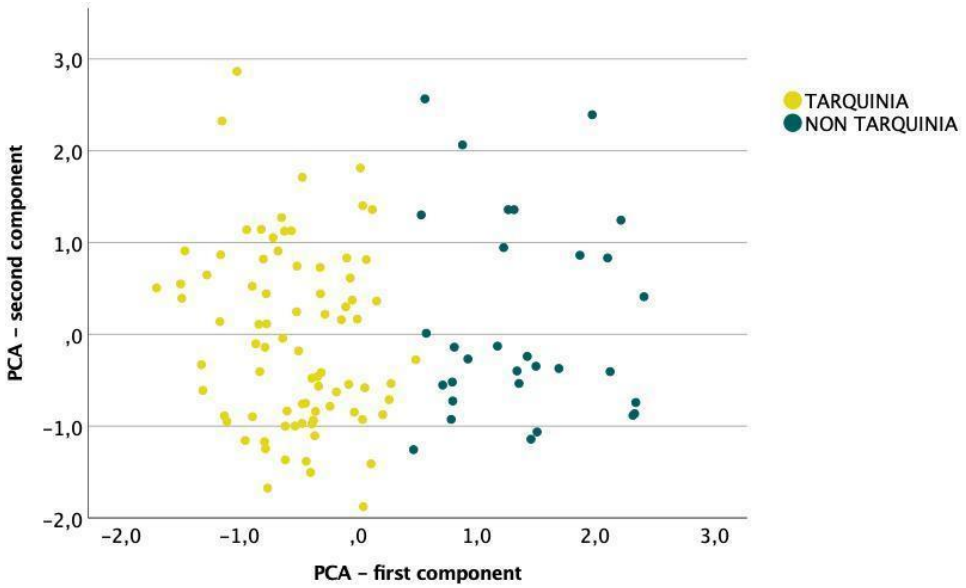


Figure 2: Data points plotted against the first two PCA components.

As far as data transformation is concerned, in some cases different scaling techniques were chosen, despite the fact that data were already normalised between 0 and 100. The more frequently chosen scaling techniques were standardisation and normalisation between 0 and 1.

Let us now focus on prediction robustness. For each considered algorithm, we inferred four different models (according to the used cross-validation technique), and we remark that these models are strictly equivalent from a statistical point of view. For this reason, they can be aggregated using majority vote as prediction and the degree of agreement as a reliability measure. We applied this new form of classification across all the 112 data points, defining as uncertain all samples for which the rate of agreement was less than one. Most of the samples were correctly classified by all of the applied methods, while for a minority of samples/measurements uncertain classification was obtained for some of the methods. Figures 3, 4 and 5 use heat maps in order to describe the samples that were uncertain for at least two learning algorithms. White and black are used for non-local and local classes, respectively. In the first column, we report the archaeological classification, and in the remaining ones we show the predictions for all learning algorithms, where the grey level accounts for the degree of agreement. This representation allows us to rapidly highlight several aspects, both on the methods and on the samples. Indeed, if we consider algorithms, those with the best/worst classification ability are pointed out by the different shadows in the relative column: for instance, for type 1 experiments (Figure 3), MLP is the worst algorithm for the classification of local materials. It is evident that LR and MLP are the worst performing algorithms. Focusing instead on samples, we can note that, if we exclude LR and MLP methods, the uncertain samples are mostly of non-local origin. This is reasonably due to the heterogeneity of the non-local class.

It is worth noting that in the three types of experiments, the uncertain samples are almost the same. In some cases, the misclassification regards all the measurements on a fragment: this indicates the requirement for a further archaeological check to exclude an error in class labelling. For instance, sample A10-3bis is uncertain for both its measurements (8.1/A10-3bis and 8.2/_2 in the figures) according to either type 1, type 2 or type 3 experiments. In some other cases, the misclassification regards only one of the measurements on the fragment, such as for one of the measurements of sample 72, which is uncertain for all the experiments. This can suggest the non-suitability of the single analytical data, possibly due to non-homogeneity of the samples itself (inclusions, decorations, alteration due to burial) or even to an error in calculating elemental concentration from the original spectrum.

The only models readily interpretable are DTs and RFs. Focusing on decision trees, the best performing four models (corresponding to the four folds of the external cross validation) developed for the three types of experiments were quite different from each other. In particular (see Figure 1S in supplementary materials) some models were very simple, since they consisted of only one decision rule based on one variable; the other models had a greater depth, that is, they contained several decision rules, which in turn were based on several variables. Reminding that the four best performing trees are equivalent from a statistical point of view, according to the shortest explanation principle (Occam's razor) we can say that the simplest models may be preferable as an operational decision support tool.

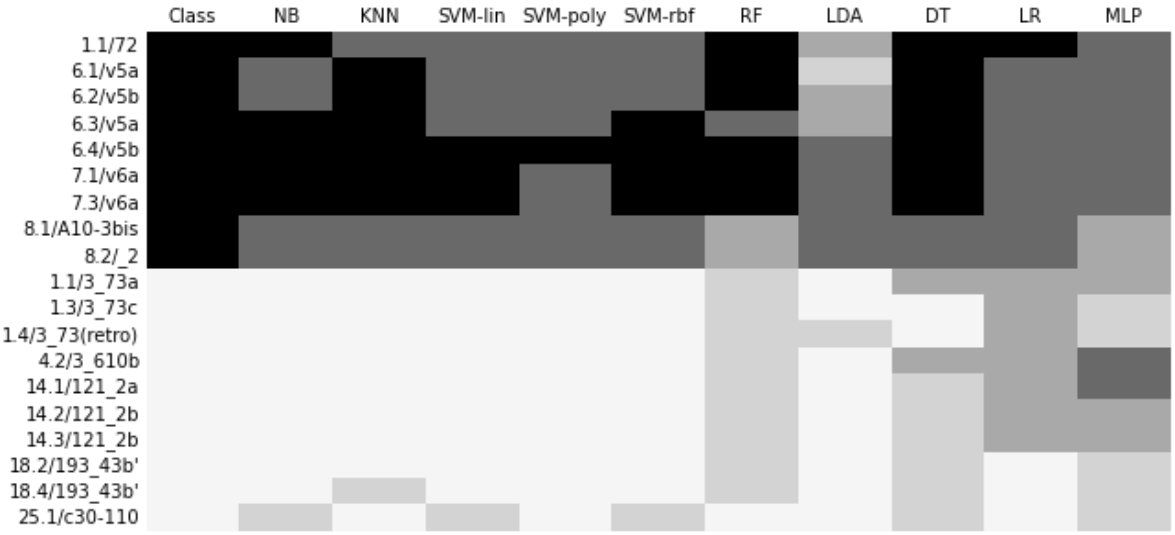


Figure 3. Classification of uncertain samples for type 1 experiments: majority on the external 4 folds, colour shade corresponds to degree of agreement, white and black are used for non-local and local classes, respectively. In the first column. Rows identifiers must be read in the following way: <fragment number> . <measure number> / <fragment>, where <fragment number> and <measure number> are conventional codes and <fragment> is the physical fragment according to Table 1.

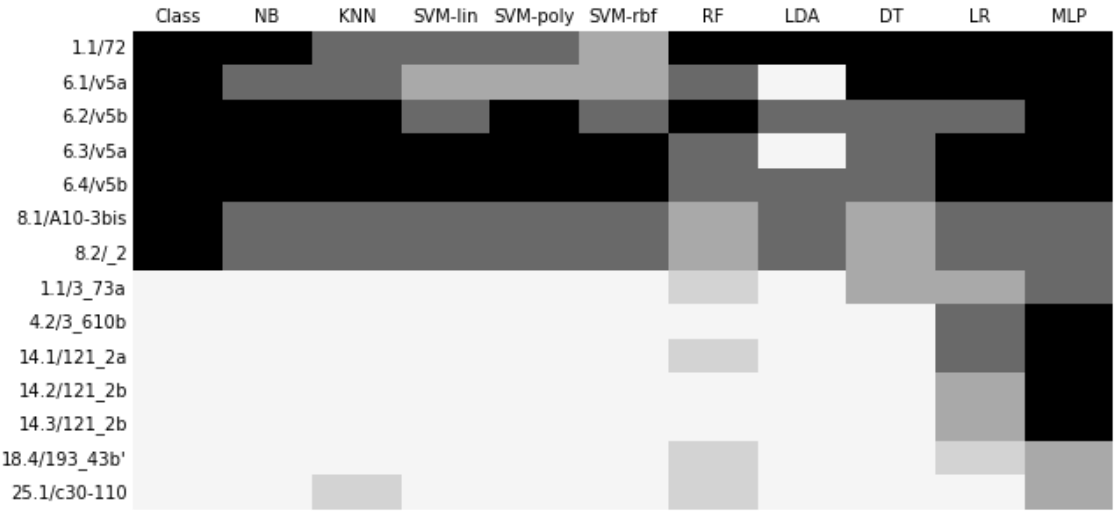


Figure 4. Classification of uncertain samples for type 2 experiments: majority on the external 5 folds, colour shade corresponds to degree of agreement, white and black are used for non-local and local classes, respectively. Rows identifiers follow the same notation as figure 3.

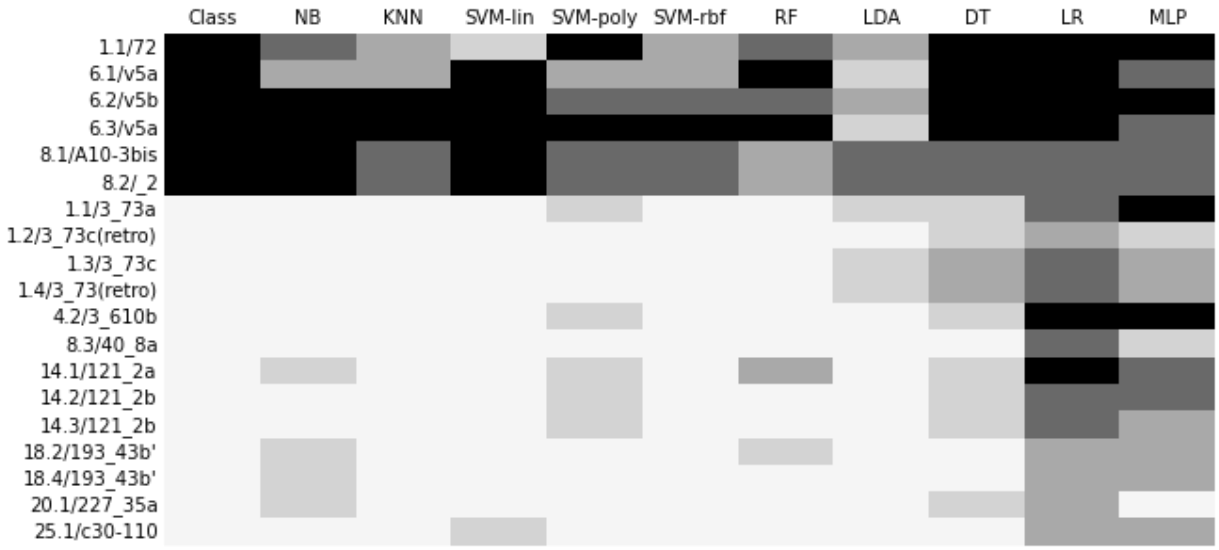


Figure 5. Classification of uncertain samples for type 3 experiments: majority on the external 5 folds, colour shade corresponds to degree of agreement, white and black are used for non-local and local classes, respectively. Rows identifiers follow the same notation as figure 3.

Conclusions

To show the potential of supervised machine learning methods, we compared the performance of ten different supervised learning algorithms for classification on a dataset of limited size, non-ideal for statistical elaborations, but typical of the archaeological context. The 36 fragments included in our set, for a total of 112 data points deriving from multiple acquisitions on each fragment, had been previously classified on archaeological basis in two main groups. The former includes samples produced in the Etruscan town of Tarquinia (local samples), the latter contains samples imported from the colony of Velia, together with other non-local fragments. The asymmetric knowledge about provenance of the two subsets, together with the limited size of the whole dataset, has been a challenge for the statistical elaboration, but it reflects the complex situation which is typically faced when dealing with ceramics finds. We planned and executed three types of experiments: in type 1 we used the complete dataset considering each data as independent, while in type 2 we considered the data from each fragment as related. In type 3, for each fragment only two data were randomly selected. To assess the reliability of our results, we considered the accuracy, the sensitivity, the specificity and the F1 score. (i.e., the harmonic mean between sensitivity and precision). As a general trend, non-linear models perform better than linear ones in almost all settings, but neural networks are always among the less performing models. We hypothesise that this is due to the high complexity of such models, which easily leads to overfitting. Indeed, very simple models such as NB and KNN consistently ranked as best across all experiments. Most of the samples were correctly classified by all of the applied methods, while for a minority of samples/measurements uncertain classification was obtained for some of the methods. The use of heat maps to describe incorrect classification allowed to highlight in a simple way whether the misclassification regards all the measurements or only one of the measurements on the fragment. In the former case, the indication is clear for the requirement for a further archaeological check to exclude an error in class labelling, while in the latter the non-suitability of the single analytical data is suggested.

The obtained results prove that Machine Learning can be of great help for archaeological classification on the basis of chemical analyses, providing a reliable and schematic picture of archaeological data even when the dataset is not suitable, in theory, for supervised learning algorithm elaboration. This approach opens the way to the building of a robust decision support system for the classification of objects whose labels are actually unknown, with the aim to confirm a supposed provenance of objects. In this perspective, as future work, one-class classification algorithms and clustering techniques are interesting learning methodologies to be considered.

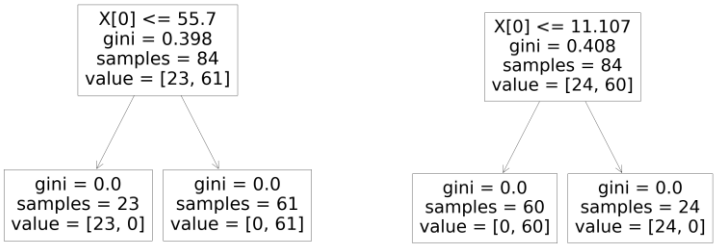
Appendix A. Supplementary materials

The software implementing our experiments and data matrix are available for replicability/reproducibility purposes at https://github.com/dariomalchiodi/JAS-Tarquinia-classification.

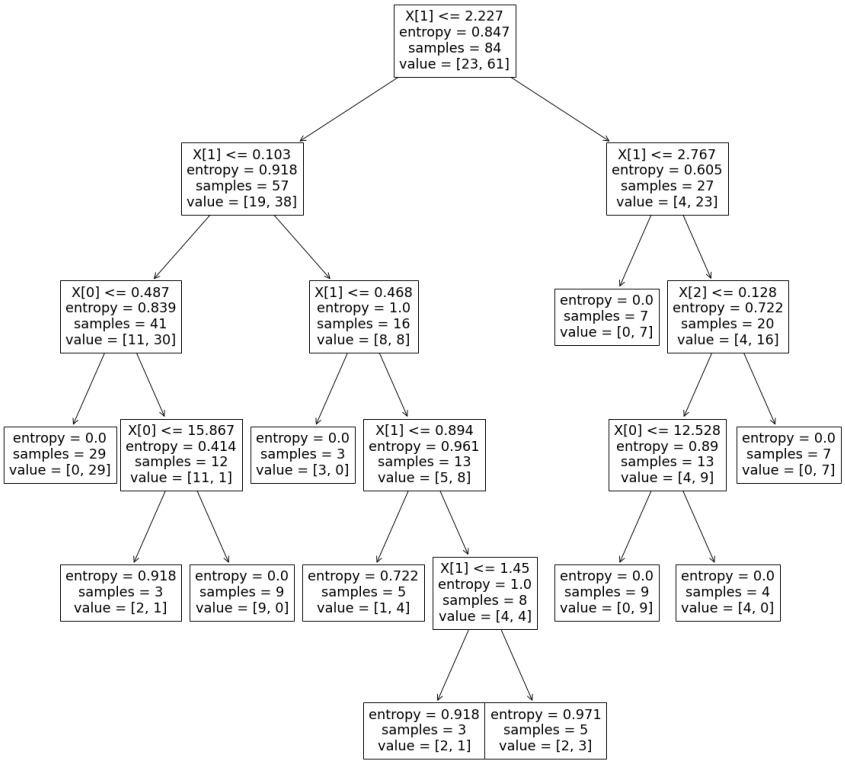
Figure 1S The four best performing decision trees developed in the four internal folds corresponding to the best mean accuracy for (a) type 1 experiments, (b) type 2 experiments and (c) type 3 experiments. The number of rules in each decision tree is the number of paths from the root to the leaves. A given variable can be considered at different decision points in the tree. Variables are ordered components in the transformed space (through PCA or Truncated SVD): X[0] is the first component, X[1] the second one and so on. At each decision point more information about the learning process is shown, namely the number of samples considered for building the rule, the mean heterogeneity of the child nodes (according to the selected heterogeneity index, Gini or entropy), and the number of data points flowing in each child node.

(a) Type 1 experiments.

a.1 The two simplest trees (depth =1). They involve only one variable, namely the first component X0, in just one rule.

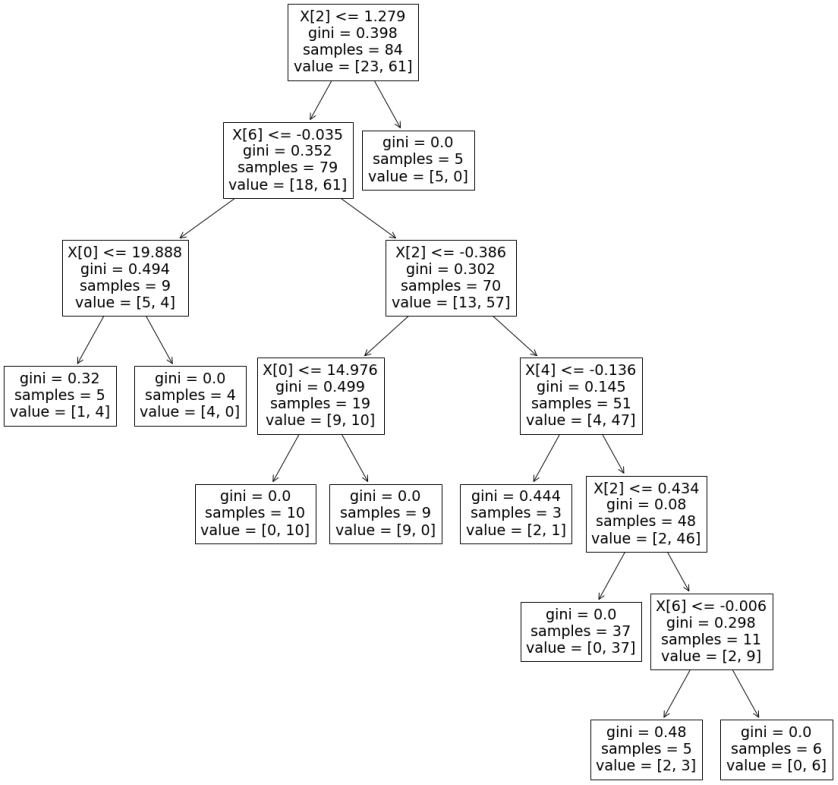


a.2 Rules are longer and involve the first three components, namely X0, X1 and X2.


X[1] <= 2.227
entropy = 0.847
samples = 84
value = [23, 61]
X[1] <= 0.103
entropy = 0.918
samples = 57
value = [19, 38]
X[1] <= 2.767
entropy = 0.605
samples = 27
value = [4, 23]
X[0] <= 0.487
entropy = 0.839
samples = 41
value = [11, 30]
X[1] <= 0.468
entropy = 1.0
samples = 16
value = [8, 8]
entropy = 0.0
samples = 7
value = [0, 7]
X[2] <= 0.128
entropy = 0.722
samples = 20
value = [4, 16]
entropy = 0.0
samples = 29
value = [0, 29]
X[0] <= 15.867
entropy = 0.414
samples = 12
value = [11, 1]
entropy = 0.0
samples = 3
value = [3, 0]
X[1] <= 0.894
entropy = 0.961
samples = 13
value = [5, 8]
X[0] <= 12.528
entropy = 0.89
samples = 13
value = [4, 9]
entropy = 0.0
samples = 7
value = [0, 7]
entropy = 0.918
samples = 3
value = [2, 1]
entropy = 0.0
samples = 9
value = [9, 0]
entropy = 0.722
samples = 5
value = [1, 4]
X[1] <= 1.45
entropy = 1.0
samples = 8
value = [4, 4]
entropy = 0.0
samples = 9
value = [0, 9]
entropy = 0.0
samples = 4
value = [4, 0]
entropy = 0.918
samples = 3
value = [2, 1]
entropy = 0.971
samples = 5
value = [2, 3]

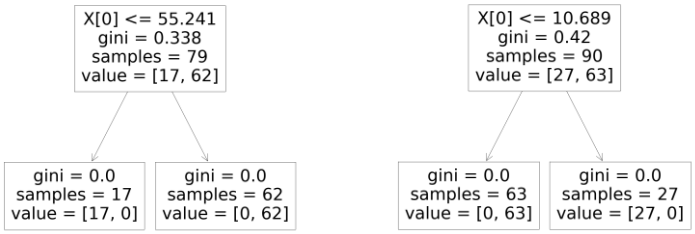
429

a.3 Rules involve many different variables.

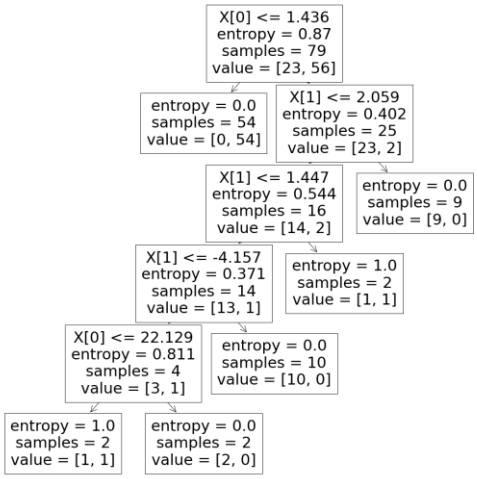


(b) Type 2 experiments.

b.1 The two simplest trees (depth =1). They involve only one variable, namely the first principal component X0, in just one rule.



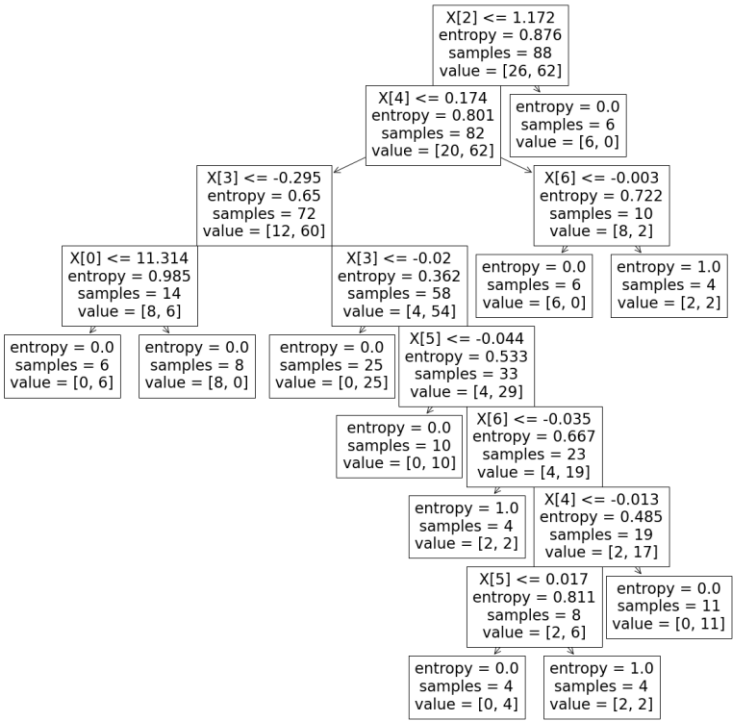
b.2 Rules are longer (depth=5) and involve only the first two components, namely X0 and X1.



438

439

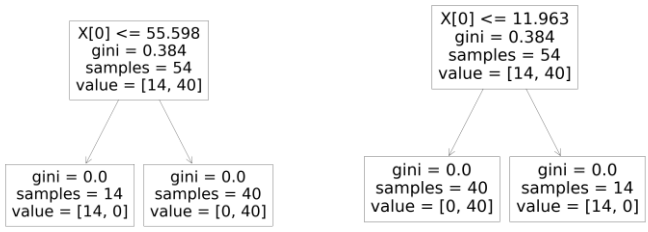
440 b.3 Rules involve many different variables.



441

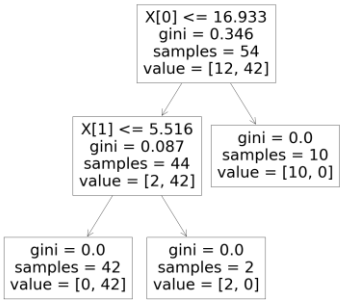
442 (c) Type 3 experiments.

443 c.1 The two simplest trees (depth =1). They involve only one variable, namely the first component X0, in just one rule.



444

445 c.2 Rules are quite short (depth=2) and involve only the first two components, namely X0 and X1.



446

c.3 Rules are longer and involve many different variables.

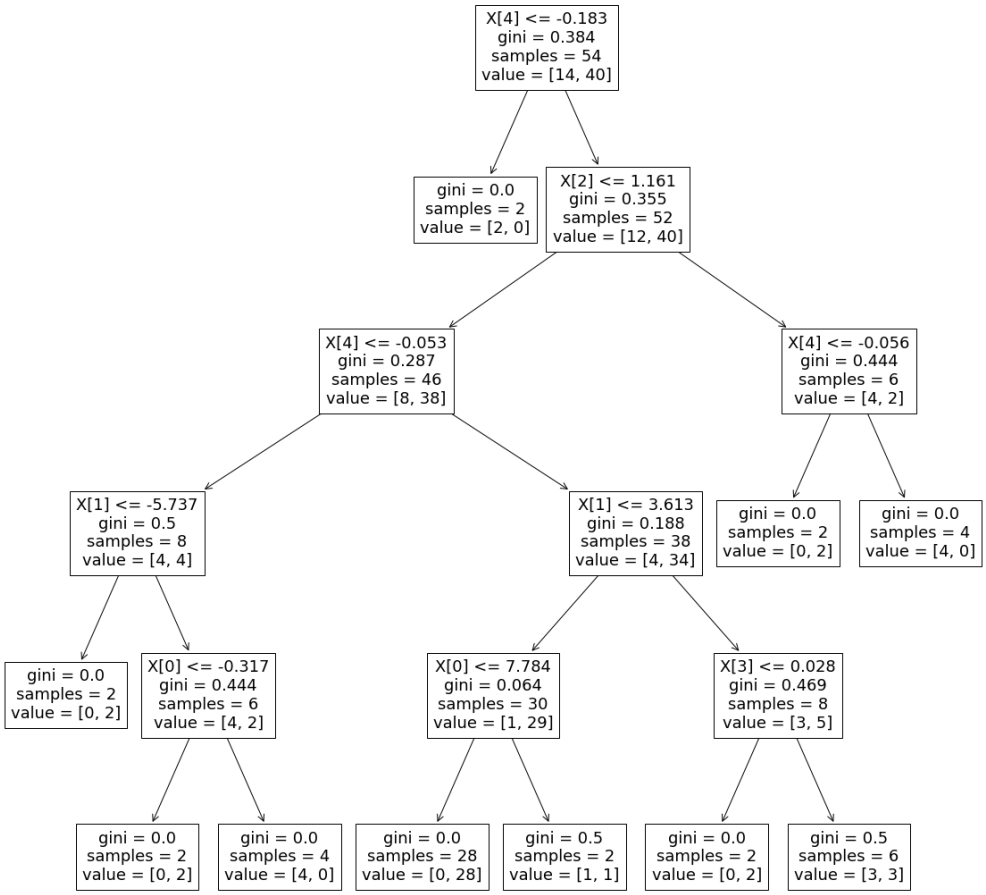
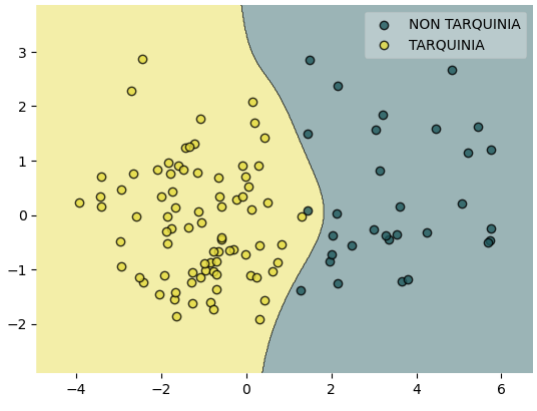
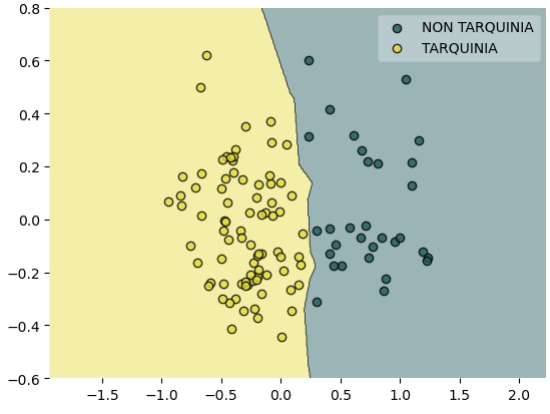


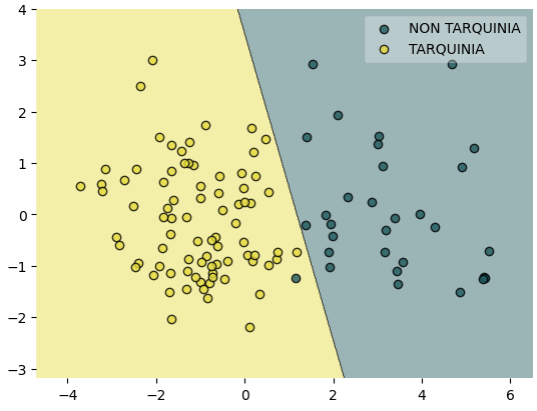
Figure 2S: For each experiment type, the graphs here below show how the best performing models relying on two extracted components separate the overall dataset.



type1, svm-poly



type2, KNN



type3, SVM-lin

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contribution

Conceptualization, LB, AZ, GR, DM; Methodology, AZ, DM, GR; Software, AZ, DM; Formal Analysis, AZ, DM; Investigation, LB, GR; Data Curation, AZ, DM; Writing – Original Draft Preparation, LB, AZ, DM, GR; Writing – Review & Editing, LB, AZ, DM, GR; Supervision, LB

References

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., Aljaaf, A.J., 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science, in: Berry, M.W., Mohamed, A., Yap, B.W. (Eds.), Supervised and Unsupervised Learning for Data Science, Unsupervised and Semi-Supervised Learning. Springer International Publishing, Cham, pp. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

Amadori, M.L., Del Vais, C., Fermo, P., Pallante, P., 2017. Archaeometric researches on the provenance of Mediterranean Archaic Phoenician and Punic pottery. Environ. Sci. Pollut. Res. 24, 13921–13949. https://doi.org/10.1007/s11356-016-7065-7

Anglisano, A., Casas, L., Anglisano, M., Queralt, I., 2020. Applications of Supervised Machine Learning Methods for Attesting Provence in Catalan Traditional Pottery Industry, Minerals 10, 8; https://doi.org/10.3390/min10010008

Aruga, R., 1998. Closure of analytical chemical data and multivariate classification. Talanta 47, 1053–1061. https://doi.org/10.1016/S0039-9140(98)00126-X

Ballirano, P., De Vito, C., Medeghini, L., Mignardi, S., Ferrini, V., Matthia, P., Bersani, D., Lottici, P.P., 2014. A combined use of optical microscopy, X-ray powder diffraction and micro-Raman spectroscopy for the characterization of ancient ceramic from Ebla (Syria) Ceramics International 40, Part B, 16409-16419

Baxter, M.J., 2006. A Review of Supervised and Unsupervised Pattern Recognition in Archaeometry*. Archaeometry 48, 671–694. https://doi.org/10.1111/j.1475-4754.2006.00280.x

Bellanti, F., Tomassetti, M., Visco, G., Campanella, L., 2008. A chemometric approach to the historical and geographical characterisation of different terracotta finds Microchem. J., 88, 113.

Bishop, R. L., Blackman, M. J., 2002. Instrumental Neutron Activation Analysis of Archaeological Ceramics: Scale and Interpretation Acc. Chem. Res. 35, 603-610.

Bonizzoni, L., Galli, A., Milazzo, M., 2010. XRF analysis without sampling of Etruscan depurata pottery for provenance classification. X-Ray Spectrom. 39, 346–352. https://doi.org/10.1002/xrs.1263

Bonizzoni, L., Galli, A., Spinolo, G., Palanza, V., 2009. EDXRF quantitative analysis of chromophore chemical elements in corundum samples. Anal. Bioanal. Chem. 395, 2021–2027. https://doi.org/10.1007/s00216-009-3158-1

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bruni, S., 2022. Etruscan Fine Ware Pottery: Near-Infrared (NIR) Spectroscopy as a Tool for the Investigation of Clay Firing Temperature and Atmosphere. Minerals 12, 412. https://doi.org/10.3390/min12040412

Bruni, S., Cariati, F., Bagnasco Gianni, G., Bonchi Jovino, M., 2001. Spectroscopic Characterization of Etruscan Depurata and Impasto Pottery from the Excavation at Pian di Civita in Tarquinia (Italy): A Comparison with Local Clay', in: Archaeology and Clays, British Archaeological Reports. I.C. Druc, Oxford.

Bruno, P., Caselli, M., Curri, M.L., Genga, A., Striccoli, R., Traini, A., 2000. Chemical characterisation of ancient pottery from south of Italy by Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES): Statistical multivariate analysis of data. Anal. Chim. Acta 410, 193–202. https://doi.org/10.1016/S0003-2670(00)00734-0

Cariati, F. Fermo, P. Gilardoni, S. Galli, A. Milazzo, M. 2003. A new approach for archaeological ceramics analysis using total reflection X-ray fluorescence spectrometry. Spectrochim. Acta, B58, 177-184.

Charalambous, E., Dikomitou-Eliadou, M., Milis, G.M., Mitsis, G., Eliades, D.G., 2016. An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships. J. Archaeol. Sci. Rep. 7, 465–471. https://doi.org/10.1016/j.jasrep.2015.08.010

Cuomo di Caprio, N., 2017. Ceramics in Archaeology. From Prehistoric to Medieval times in Europe and the Mediterranean: Ancient Craftsmanship and Modern Laboratory Techniques. L'Erma di Bretschneider

Descantes, C., Neff, H., Glascock, M.D., Dickinson, W.R., 2001, Chemical characterization of Micronesian ceramics through instrumental neutron activation analysis: A preliminary provenance study J. Archaeol. Sci., 28, 1185.

Fermo, P., Andreoli, M., Bonizzoni, L., Fantauzzi, M., Giubertoni, G., Ludwig, N., Rossi, A., 2016. Characterisation of Roman and Byzantine glasses from the surroundings of Thugga (Tunisia): Raw materials and colours. Microchem. J. 129, 5–15. https://doi.org/10.1016/j.microc.2016.05.014

Fermo, P., Cariati, F., Ballabio, D., Consonni, V., Bagnasco Gianni, G., 2004. Classification of ancient Etruscan ceramics using statistical multivariate analysis of data. Appl. Phys. A 79, 299–307. https://doi.org/10.1007/s00339-004-2520-6

Fermo, P., Delnevo, E., Lasagni, M., Polla, S., de Vos, M., 2008. Application of chemical and chemometric analytical techniques to the study of ancient ceramics from Dougga (Tunisia). Microchem. J., SELECTED PAPERS FROM THE 1st INTERNATIONAL

SYMPOSIUM ON MULTIVARIATE ANALYSIS AND CHEMOMETRICS FOR CULTURAL HERITAGE AND ENVIRONMENT Nemi, Italy 2 - 4 October 2006 88, 150–159. https://doi.org/10.1016/j.microc.2007.11.012

Frahm, E., 2018. Ceramic studies using portable XRF: From experimental tempered ceramics to imports and imitations at Tell Mozan, Syria. J. Archaeol. Sci. 90, 12–38. https://doi.org/10.1016/j.jas.2017.12.002

Freitas, R.P., Coelho, F.A., Felix, V.S., Pereira, M.O., de Souza, M.A.T., Anjos, M.J., 2018. Analysis of 19th century ceramic fragments excavated from Pirenópolis (Goiás, Brazil) using FT-IR, Raman, XRF and SEM. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 193, 432–439. https://doi.org/10.1016/j.saa.2017.12.047

Galli, A., Bonizzoni, L., Sibilia, E., Martini, M., 2011. EDXRF analysis of metal artefacts from the grave goods of the Royal Tomb 14 of Sipán, Peru. X-Ray Spectrom. 40, 74–78. https://doi.org/10.1002/xrs.1298

Hazenfratz, R., Munita, C.S., Neves, E.G., 2017. Neural Networks (SOM) Applied to INAA Data of Chemical Elements in Archaeological Ceramics from Central Amazon. STAR Sci. Technol. Archaeol. Res. 3, 334–340. https://doi.org/10.1080/20548923.2018.1470218

Idjouadiene, L., Mostefaoui, T.A., Djermoune, H., Bonizzoni, L., 2019. Application of X-ray fluorescence spectroscopy to provenance studies of Algerian archaeological pottery. X-Ray Spectrom. 48, 505–512. https://doi.org/10.1002/xrs.3020

Jasiewicz, J., Niedzielski, P., Krueger, M., Hildebrandt-Radke, I., Michałowski, A., 2021. Elemental variability of prehistoric ceramics from postglacial lowlands and its implications for emerging of pottery traditions – An example from the pre-Roman Iron Age. J. Archaeol. Sci. Rep. 39, 103177. https://doi.org/10.1016/j.jasrep.2021.103177

Jones, R.E., 1986. Greek and Cypriot Pottery: A Review of Scientific Studies. British School at Athens, Athens.

Kennett, D. J., Sakai, S., Neff, H., Gossett, R., Larson, D.O., 2002. Compositional Characterization of Prehistoric Ceramics: A New Approach. J. Archaeol. Sci., 29, 443-455.

Li, B.P., Zhao, J.X., Greig, A., Collerson, K.D., Feng Y.X., 2006. Characterisation of Chinese Tang sancai from Gongxian and Yaozhou kilns using ICP-MS trace element and TIMS Sr–Nd isotopic analysis, Journal of Archaeological Science 33, 56-62.

Liritzis, I., Xanthopoulou, V., Palamara, E., Papageorgiou, I., Iliopoulos, I., Zacharias, N., Vafiadou, A., Karydas, A.G., 2020. Characterization and provenance of ceramic artifacts and local clays from Late Mycenaean Kastrouli (Greece) by means of p-XRF screening and statistical analysis. J. Cult. Herit. 46, 61–81. https://doi.org/10.1016/j.culher.2020.06.004

Maritz, J.S., Lwin, T., 2018. Empirical Bayes Methods, 1° edizione. ed. Routledge, Place of publication not identified.

Neff, H., 2000. Neutron activation analysis for provenance determination in archaeology, in: Modern Analytical Methods in Art and Archaeology, Chemical Analysis. E. Ciliberto and G. Spoto.

Padilla, R., Espen, P.V., Torres, P.P.G., 2006. The suitability of XRF analysis for compositional classification of archaeological ceramic fabric: A comparison with a previous NAA study. Anal. Chim. Acta 558, 283–289. https://doi.org/10.1016/j.aca.2005.10.077

Papageorgiou, I., 2020. Ceramic investigation: how to perform statistical analyses. Archaeol. Anthropol. Sci. 12, 210. https://doi.org/10.1007/s12520-020-01142-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: Machine Learning in Python. Mach. Learn. PYTHON 6.

Robertson, J. D., Neff, H., Higgins, B., 2002. Microanalysis of ceramics with PIXE and LA-ICP-MS. NIM B 189, 378-381.,

Romano, F.P., Pappalardo, G., Pappalardo, L., Garraffo, S., Gigli, R., Pautasso, A., 2006. Quantitative non-destructive determination of trace elements in archaeological pottery using a portable beam stability-controlled XRF spectrometer. X-Ray Spectrom. 35, 1–7. https://doi.org/10.1002/xrs.880

Ruschioni, G., Micheletti, F., Bonizzoni, L., Orsilli, J., Galli, A., 2022. FUXYA2020: A Low-Cost Homemade Portable EDXRF Spectrometer for Cultural Heritage Applications. Appl. Sci. 12, 1006. https://doi.org/10.3390/app12031006

Saleh, M., Bonizzoni, L., Orsilli, J., Samela, S., Gargano, M., Gallo, S., Galli, A., 2020. Application of statistical analyses for lapis lazuli stone provenance determination by XRL and XRF. Microchemical Journal 154, 104655.

Sun, H., Liu, M., Li, L., Yan, L., Zhou, Y., Feng, X., 2020. A new classification method of ancient Chinese ceramics based on machine learning and component analysis. Ceram. Int. 46, 8104–8110. https://doi.org/10.1016/j.ceramint.2019.12.037

Tite, M. S., Kilikoglou, V., Vekinis, G., 2003. Strength, Toughness and Thermal Shock Resistance of Ancient Ceramics, and Their Influence On Technological Choice. Archaeometry 43, 301-324.

Todorov, V., R package rrcov: Scalable Robust Estimators with High Breakdown Point, https://CRAN.R-project.org/package=rrcov, 2022

Sciau, P., Goudeau, P., 2015. Ceramics in art and archaeology: a review of the materials science aspects. Eur. Phys. J. B 88, 132. https://doi.org/10.1140/epjb/e2015-60253-8

Veneranda, M., Prieto-Taboada, N., Costantini, I., Francesco, A.M.D., Castro, K., Madariaga, J.M., Arana, G., 2022 Portable XRF and LIBS combined with chemometrics: a novel method for the in-situ geochemical sourcing of obsidian artefacts.

Wagner, U., Gebhard, R., Häusler, W., Hutzelmann, T., Riederer, J., Shimada, I., Sosa, J., Wagner, F.E., 1999. Reducing firing of an early pottery making kiln at Batán Grande, Peru: A Mössbauer study. Hyperfine Interactions 122, 163–170.