CSCI 6443

Data Mining

Term Paper

Title: **Unearthing Hidden Insights: Data Mining in Archaeology**

Submitted by: **Harshita Chadha (GWID – G40737617)**

Abstract

From social media posts to the Internet of Things (IoT) sensors, the contemporary world generates millions of terabytes of data each day. However, obscured behind this influx of modern-day big data, is the vast sea of unutilized information from history waiting to be analyzed to understand the evolution of society as we know it. Fortunately, a paradigm shift is now underway and many initiatives such as the digitization of ancient manuscripts, etc are underway to establish open-access archaeological data repositories. The immense wealth of data accumulated over millennia is now becoming accessible, opening doors to a new frontier in archaeological research. Amidst this newfound availability of data, mining, and analysis techniques are increasingly being applied to uncover interesting insights, discern temporal trends, etc. This newly emerging field of archaeological data mining serves as a bridge between the wisdom of antiquity and the data-driven age, enriching our understanding of human history and culture in ways that were once unimaginable.

The objective of this term paper is to explore the impact of data mining techniques in archaeological research enhancement. To this effect, a detailed study of the various techniques and latest research presentations that utilize data mining to derive previously unknown insights from archaeological data was undertaken. Moreover, to demonstrate the applicability of data mining in the field of archaeology, three case studies are also presented.

This term paper examines the use of data mining techniques in archaeology, examining different approaches and how they may be applied to uncover insights from archaeological data that were not previously known. The content delves into collaborative aspects of interdisciplinary research and attempts to shed light on the broader implications of utilizing data mining in the field of archaeology, a practice that contributes to a nuanced understanding of societal structures and cultural exchange throughout history. The contents of this article aim to capture the ongoing discourse of gaining insights into the past through the lenses of contemporary analytical frameworks as archaeological research ventures into new and exciting territories.

Table of contents

1. Introduction

It can be argued that one of the main aims of the field of psychology is to conduct investigations into gathered residual evidence in an attempt to understand social and societal contexts of the past. When one thinks of archaeological excursions and information gathering exercises, computers and databases are far from our mind. Instead, the image that comes to mind is that of a group of people wearing hats kneeling with hand tools in the middle of sand filled desserts, carefully sifting away to dig-up earthly goods from pat. Archaeology traditionally has been a computer shy field but in recent years due to advent of advanced techniques many headways have been made that have had the professionals in the field relying more on computers for sound results [2-3]. From uncovering hidden Mayan civilizations through the application of lidar technology [4] to digitally reconstructing the face of King Tutankhamun using advanced computing methods [5], modern technology has fundamentally transformed the landscape of archaeology, offering unprecedented insights and reshaping our understanding of ancient civilizations.

In terms of statistical and data based analysis, techniques like clustering have been used in the field of archaeology since the 1970s. An example of this is the work of Hodder et al [6]. Their book on spatial analysis in archaeology has emerged as a pioneering study in the field of application of modern statistical and quantitative techniques to archaeology. However, the use of advanced data mining tools had been few and far between in the field up until a few years ago. This lack of research can be attributed to the lack of availability of properly documented data sources that lie at the centre of data mining analysis.

Each year, billions of dollars are spent by public and government funded private agencies to carry out archaeological research [7]. However, despite the abundance of social context data collected over many years, preservation and accessibility remain challenges. Long-range open-source data, crucial for understanding phenomena like the loss of identity in migrant populations, often lacks proper preservation. In the US, approximately 30,000 mandated archaeological studies annually yield valuable results, but their storage in outdated formats hinders practical use [8]. Fortunately, a paradigm shift is now underway. Initiatives such as York University's Archaeological Data Service [9] and Oxford University's ORAU [10] are now addressing these issues by preserving and providing accessible guidelines for sustainable data storage in archaeology.

Today, archaeology is rapidly embracing 'digital humanities,' leveraging well-documented datasets and advanced analytics tools to revolutionize the field [11]. While archaeological datasets may not rival those in core computational fields, the utilization of previously untapped resources is reshaping research paradigms. Archives are increasingly digitized, and machine learning and data mining uncover valuable insights. Traditional closed reading methods are giving way to quantitative approaches, enabling a comprehensive perspective and revealing previously imperceptible trends in findings [12].

In line with the present line of argumentation, this article explores the use of data mining techniques in archaeology during present times, investigating diverse approaches to reveal new insights from archaeological data. The article is divided into five main sections. The first and present section serves as an introduction into the theme of the article. The second section presents a detailed literature survey that explores past work done in the field with an aim to capture the ongoing dialogue on understanding the past through contemporary analytical frameworks, marking the exciting frontiers of archaeological research. Following this, three case studies are presented to demonstrate the applicability of data mining principles in the field of archaeology.

The first case study utilizes the Beazley Archive Pottery Database (BAPD) about ancient Greek artefacts with the aim to understand the evolution of trends in pottery design over time and what might these trends indicate in terms of cultural and socio-political shifts within the society of ancient Greece itself. The second case study utilizes the Southern African Radiocarbon Dating Database (SARD) to create a Random Forest classifier to predict the Archaeological time periods to which an unearthed artefact might belong to in the African context. The third case study utilizes the Digital Archive for Grave Goods: Objects and Death in Later Prehistoric Britain to explore relationships amongst the burial site attributes with an aim to identify the significance of specific goods occurrences within distinct cultural contexts. The conclusions of the undertaking and some challenges facing the field in the future are also discussed. In the section that follows, the important terms and some abbreviations used throughout the rest of the article are tabulated.

1.1.Abbreviations and Important Terms

Table 1 below, presents a comprehensive list of the important terms and abbreviations that are referenced throughout the remaining text.

Table 1: Abbreviations and Terms

Term / Abbreviation	Definition
lidar	Light Detection and Ranging. It is a remote sensing method used to examine the surface of the Earth.
ORAU	Oxford Radiocarbon Accelerator Unit
BAPD	Beazley Archive Pottery Database. Used for case study number 1.
SARD	Southern African Radiocarbon Dating Database. Used for case study number 2.
GIS	Geographical Information System
PCA	Principal Component Analysis [38]
KCA	Kernel Density Estimation Clustering Algorithm [39]
K-Means	A clustering algorithm [40]
DBSCAN	Density-based spatial clustering of applications with noise [41]
C4.5 Algorithm	A decision tree generation algorithm [42]
Euclidean Distance	Distance calculation mechanism [43]
VGG-16	A convolutional neural network architecture [44]
KNN	K Nearest Neighbour [45]
SVM	Support Vector Machine [46]
SSVM	Smooth Support Vector Machine [47]
Gabor Wavelet Transformations	Complex functions constructed to serve as a basis for Fourier transforms [48]
SIMCA	Soft independent modelling of class analogy [49]
LDA	Linear Discriminant Analysis [50]

Learning Vector Quantizer	Artificial neural network algorithm that gives the optimal training instances [51]
fuzzywuzzy library	Python string matching library [52]
Bagging	Bootstrap Aggregation [53]
csv file	comma separated files
Apriori Algorithm	Association rule mining algorithm [54]

2. Literature Survey: Past Related Work

As mentioned in section 1, one of the earliest known examples of the use of modern data mining tools in archaeology can be seen in the work of Hodder et al [6]. The authors through their study show how various clustering techniques, when sensitively employed, can dramatically extend and refine the information presented in distribution maps and other analyses of spatial relationships. Clustering analysis remains, to this date, one of the most widely used data mining technique in the field of archaeology and clustering capabilities are built into modern day GISs which are indispensable to the field of archaeology. GIS stands for geographical information system and is used in archaeology to study the spatial distribution of artefacts. Coupled with strong clustering capabilities, this technology can help unveil settling patterns, artefact deposition patterns, highlight activity hubs, and facilitate chronological and cultural phase analysis. An example of such a system is ArGIS [13].

A study that uses a combination of different data mining techniques to investigate found artefacts is the one conducted by Fermo et al [14]. In their work, they examined archaeological ceramic shards belonging to three principal but distinct classes namely the African Red Slip Ware, the Dougga ware, and the African Cooking ware from modern day Tunisia using various different classes of mining techniques. They utilized algorithms such as PCA for pattern recognition, KNN and SIMCA for classification and hierarchical clustering in their study to unveil interesting patterns. Next, Sanchez-Romero et al [15] explore spatial analysis methods in Palaeolithic site studies, assessing georeferencing, spatial modelling, density analyses, hotspots, and unsupervised classification. Emphasizing interdisciplinary approaches, the paper contrasts clustering algorithms like Kernel Density Estimation Clustering Algorithm (KCA) and k-means for enhanced spatial archaeological insights.

In terms of novel archaeology specific algorithms, Casper et al [16] propose Archsphere, a clustering algorithm that is designed specifically keeping in mind the nature of archaeological data. Their proposed approach is positioned to work better than existing algorithms and accounts for shortfalls in generic cluster algorithms like the difficulty to cluster point clouds with varying densities in DBSCAN or the absence of a notion of noise in k-means. Their algorithm detects clusters with varying densities, incorporating a structural parameter and spatial location information, represented by connected spheres around each point, where two objects are considered connected if their spheres touch or overlap; clusters are automatically identified using Breadth First Search (BFS), with parameters for minimum cluster points, weight, and noise, and distances are determined based on sightlines between monuments in three dimensions.

In order to test the hypothesis of the existence of nine spatially distinct local residence groups exiting within Lualualei archaeological records of Hawaii, Dixon et al [17] employed K-means clustering analysis. The K-means analysis assumed nine clusters based on the initial grouping of sites, with the goal of minimizing variability within clusters and maximizing variability between clusters. The results indicated the presence of eight clusters, interpreted as local residence groups, revealing spatial patterns and relationships within the archaeological data. The study combined the K-means analysis with a rank-size analysis of permanent habitation and ritual structures to further explore socio-political centralization in Lualualei valley, providing insights into settlement patterns and land use practices.

Motivated to define a more suitable method for assigning samples to groups in archaeological materials, Lopez-Garcia et al [18] presented a comparison between three non-supervised model-based clustering methods focusing on the selection of informative variables and assigning samples to groups. Once the important variables for clustering have been selected, a data projection method called PSwarm [19] is used for the projection of high-dimensional data.

Bi et al [20] worked with spatial data from the Jiangzhai site in what is modern day China. They worked with the relic distribution map of this region and transformed the images into vector graphics and used these representations for mining information. The study uses decision tree classification with the C4.5 algorithm and k-means clustering algorithm to analyze the distribution rules of house groups and internal structure of the Jiangzhai site. Further, Rasheed et al [21], present a novel framework to solve the problem of classification and reconstruction of archaeological fragments. The proposed methodology consists of two phases: Classification of Ancient Fragments (CAF), and Reconstruction of Ancient Objects (RAO). Classification is performed based on texture and color properties extracted from images using a custom Euclidean distance based approach. The proposed method achieved a success rate higher than previous studies when applying the same test dataset.

A very interesting application of classification in the archaeological domain is the work presented by Canul-Ku et al [22]. They attempt to classify artefacts based on their three dimensional representations by generating 3-Dimensional shape descriptors using the VGG-16 neural network. Once generated, these vectors were fed to classification algorithms such as KNN and SVM/SSVM to predict the actual shape of the artefact. In terms of final outcomes, while the performance of their KNN and SVM/SSVM was comparable and none had more advantage in terms of actual accuracy metrics, because the KNN took greater time to classify higher dimension neighbours and was more prone to overfitting, they preferred SVM/SSM.

Markidis et al [23] present a methodology for the automatic classification of archaeological sherds. Sherds are fragments of relics with little to no marking on them making manual classification and categorization difficult. They used features on the back and front of the shards such as color, texture, chrominance etc and converted that into global descriptor vectors and then used KNN to classify the sherds into one of the ground truth classes based on previously classified samples. On the other hand, Li-Ying et al [24] classified ancient sherds using solely texture-based features, which were extracted by applying Gabor wavelet transformations. These features were then used to classify sherds using an unsupervised kernel fuzzy clustering algorithm [25].

Archaeometry data provides information about the chemical composition of the found artifact [26]. An early attempt to apply classification to archaeometry data was done by Kowalski et al [27]. They used obsidian samples from around the region of northern California and used clustering algorithms such as ISODATA to achieve hyperplane separation amongst the data points and hence attempted to classify them into groups of similar artefacts. They also used the k-nearest neighbour algorithm and experimented with different values of K to observe the evolution of produced clusters. Mussumarra et al [28] too worked with archaeometry data but instead of working with ceramics, they attempted to use the percentage of acid soluble components (ASC) and the aggregate granulometric distribution in mortars from two classes – Gothic and Flemish wall painting plaster samples. For classification, they used the SIMCA binary classification algorithm.

Further, Baxter [29] conducted a study applying data mining on Israeli glass data and analyzed 241 specimens with techniques like PCA, LDA, hierarchical clustering, k-means, k-medoids, fuzzy clustering, KNN, logistic regression, SVM, and decision trees based on Gini index. The glass specimens, based on seven chemical variables, were classified into five groups linked to specific sites and furnaces by the original researchers. Garcia-Heras et al [30] used chemical characteristics of archaeological ceramics and employed a PCA based clustering for identification of distinct groups. In their article, Charalambous et al [31] utilize classification techniques such as KNN classifier, C4.5 decision tree algorithm and a neural network based Learning Vector Quantizer to classify a dataset of 177 ceramics obtained from early to middle bronze age period from Cyprus. Their work reinforces the importance and support that data mining offers to archaeologists by helping classify ceramics which could not be classified into categories based on traditional ceramic petrography based approaches.

Moving on from Archaeometry to the problem of aerial photography classification, Kobylinski et al [32] employed association rule mining and classification using the EdgeFlow algorithm [33] for image segmentation. They worked to extract color and texture features from segmented images, normalized them, and went on to create a visual dictionary with this information. They built their image classifier using association rule mining on a learning set to identify key relationships between image features and categories. This compact classifier automates categorization by labelling new photographs based on matching rules, defaulting to a class label if no rules are met.

Furthering the exploration of the use of rule mining, Wilcke et al [34] proposed the MINOS pipeline which can be utilized to mine association rules in knowledge graphs with a particular focus on archaeological applications. Their methodology is based on the SWARM association rule mining algorithm [35] which is specifically designed for RDF or Resource Description Framework data in the context of knowledge graphs to automatically mine semantic association rules.

Another interesting study is the one conducted by Brown et al [36] that utilizes Text mining approaches such as word frequency and n-grams to mine information from set of oral histories from the anthracite coal mining region of north-eastern Pennsylvania, where the industry was dying, and communities remembered work and the struggle to survive during the industry's decline. The most common word identified was "mines," and there was also an emphasis on family indicated with discussions that included terms like "father," "mother," "family," "born," "married," and "children." There was also some evidence to suggest ethnic tensions and identities in the community, as interviewees emphasized ethnic affiliations, such as "Polish," "Irish," and "English." These terms indicated a heightened awareness of collective identity and ethnic differences.

Table 2: Overview of Surveyed Articles

Authors	Algorithms Used	Technique
Hodder et al [6]	K-means Clustering	Clustering
Fermo et al [14]	PCA, K-Nearest Neighbour, SIMCA, Hierarchical clustering	Clustering, Classification, Pattern

		Discovery
Sanchez-Romero et al [15]	the Kernel Density Estimation Clustering Algorithm (KCA) , K-Means Clustering	Clustering
Casper et al [16]	Archsphere Clustering Algorithm	Clustering
Dixon et al [17]	K-means Clustering	Clustering
Lopez-Garcia et al [18]	PSwarm Clustering	Clustering
Bi et al [20]	Decision tree and k means clustering	Clustering, Classification
Rasheed et al [21]	Euclidian distance based approach	Classification
Canul et al [22]	KNN, SVM	Classification
Markidis et al [23]	KNN	Classification
Li-Ying et al [24]	Fuzzy Clustering	Clustering, Classification
Kowalski et al [27]	ISODATA, KNN	Clustering, Classification
Mussumarra et al [28]	SIMCA binary classification, PCA	Classification
Baxter [29]	PCA, LDA, hierarchical clustering, K-Means, k-medoids, KNN, Logistic regression, SVM, ginny index based decision tree	Classification, clustering
Garcia-heras et al [30]	PCA based clustering	clustering
Charalambous et al [31]	KNN, C4.5 decision tree, neural network based Learning Vector Quantisation	classification
Wilcke et al [34]	MINOS Pipeline	Rule Mining
Brown et al [36]	Word Frequency and N-grams	Text Mining
Kobylinski et al [32]	EdgeFlow Algorithm and Association Rule Mining for Classification	Classification

In table 2 above, a brief summary of the research articles surveyed has been provided.

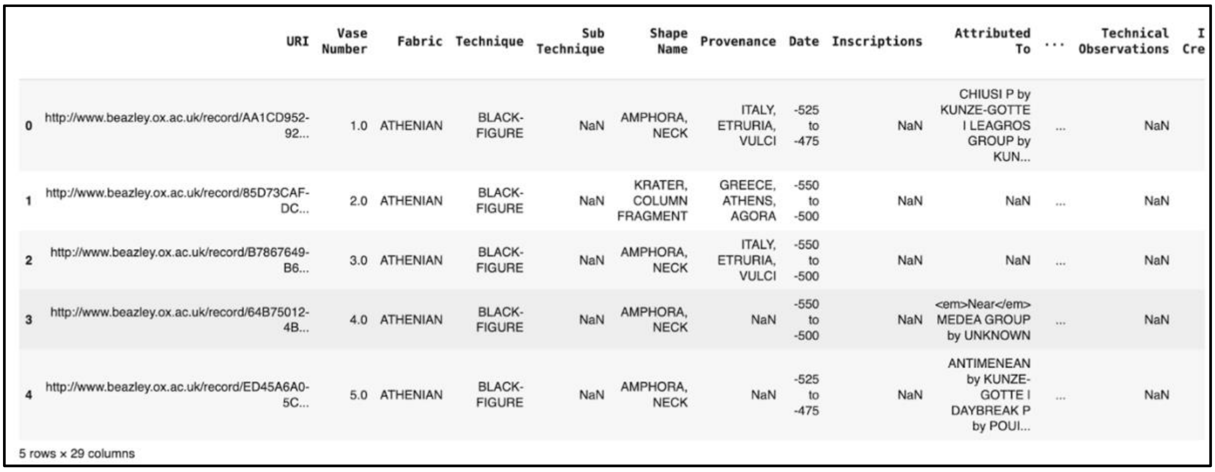
3. Applications of Data Mining in Archaeology: A Case Study Approach

In this section of the article, three case studies are presented to demonstrate the practical applicability of data mining techniques in the field of Archaeology. The first case study utilizes a dataset on ancient Greek pottery. Using the KModes algorithm, we cluster the data points across different time frames to understand the shift in the kind of pottery production patterns and practices. In the second case study, we work with a radiocarbon dating dataset to create a random forest classifier that given information about a given material can help predict the Archaeological period to which that material might belong. Finally, in the third case study, we work with a dataset about ancient graves across the United Kingdom to identify frequently buried object sets using the Apriori algorithm. A detailed description of the data sources, the pre-processing steps, the application of data mining techniques, and the consequent results and inferences drawn are presented in the sections below.

3.1. Case Study 1 – Clustering with Greek Pottery

In this case study, the KModes clustering algorithm has been applied to the data across different timelines helped identify trend groups across different geographical locations and the shift and spread in pottery techniques. The case study works with data obtained from the Beazley Archive Pottery Database (BAPD) which is the world's largest database that contains information on ancient Greek painted Potter. The BAPD is currently being maintained by the

Classical Art Research Centre at Oxford University [37]. The dataset is composed of information on ancient vases most of which are estimated to have been created during the period from 6th to 4th century BC. For the present case study, a subset of this data containing information about the vases was worked with. The figure X below illustrates an initial snapshot of the data utilised.

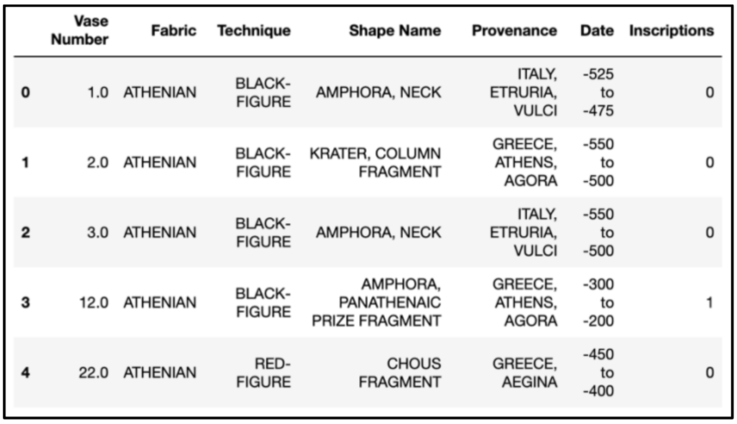


	URI	Vase Number	Fabric	Technique	Sub Technique	Shape Name	Provenance	Date	Inscriptions	Attributed To	...	Technical Observations	I Cre
0	http://www.beazley.ox.ac.uk/record/AA1CD952-92...	1.0	ATHENIAN	BLACK-FIGURE	NaN	AMPHORA, NECK	ITALY, ETRURIA, VULCI	-525 to -475	NaN	CHIUSI P by KUNZE-GOTTE I LEAGROS GROUP by KUN...	...	NaN	
1	http://www.beazley.ox.ac.uk/record/85D73CAF-DC...	2.0	ATHENIAN	BLACK-FIGURE	NaN	KRATER, COLUMN FRAGMENT	GREECE, ATHENS, AGORA	-550 to -500	NaN	NaN	...	NaN	
2	http://www.beazley.ox.ac.uk/record/B7867649-B6...	3.0	ATHENIAN	BLACK-FIGURE	NaN	AMPHORA, NECK	ITALY, ETRURIA, VULCI	-550 to -500	NaN	NaN	...	NaN	
3	http://www.beazley.ox.ac.uk/record/64B75012-4B...	4.0	ATHENIAN	BLACK-FIGURE	NaN	AMPHORA, NECK	NaN	-550 to -500	NaN	<em>Near</em> MEDEA GROUP by UNKNOWN	...	NaN	
4	http://www.beazley.ox.ac.uk/record/ED45A6A0-5C...	5.0	ATHENIAN	BLACK-FIGURE	NaN	AMPHORA, NECK	NaN	-525 to -475	NaN	ANTIMENEAN by KUNZE-GOTTE I DAYBREAK P by POUI...	...	NaN	

5 rows × 29 columns

Figure 1: Initial Snapshot Of The BAPD Dataset

The dataset in its raw form contained 29 columns with information such as the shape, of the vases, their provenance or location of origin, estimated date of creation, inscribed text, etc. In addition to such relevant columns for our present application, additional columns such as URI, LIMC ID, etc were also present that were not of use for the present application and were hence removed. Furthermore, an initial look into Null values indicated that columns such as measurement, volume, weight, etc had 99-100% of their values as null values, so these were also filtered out. This resulted in a dataset containing seven most relevant columns for the present analysis. These were – ['Vase Number', 'Fabric', 'Technique', 'Shape Name', 'Provenance', 'Date', 'Inscriptions']. This is shown in the figure below.



	Vase Number	Fabric	Technique	Shape Name	Provenance	Date	Inscriptions
0	1.0	ATHENIAN	BLACK-FIGURE	AMPHORA, NECK	ITALY, ETRURIA, VULCI	-525 to -475	0
1	2.0	ATHENIAN	BLACK-FIGURE	KRATER, COLUMN FRAGMENT	GREECE, ATHENS, AGORA	-550 to -500	0
2	3.0	ATHENIAN	BLACK-FIGURE	AMPHORA, NECK	ITALY, ETRURIA, VULCI	-550 to -500	0
3	12.0	ATHENIAN	BLACK-FIGURE	AMPHORA, PANATHENAIC PRIZE FRAGMENT	GREECE, ATHENS, AGORA	-300 to -200	1
4	22.0	ATHENIAN	RED-FIGURE	CHOUS FRAGMENT	GREECE, AEGINA	-450 to -400	0

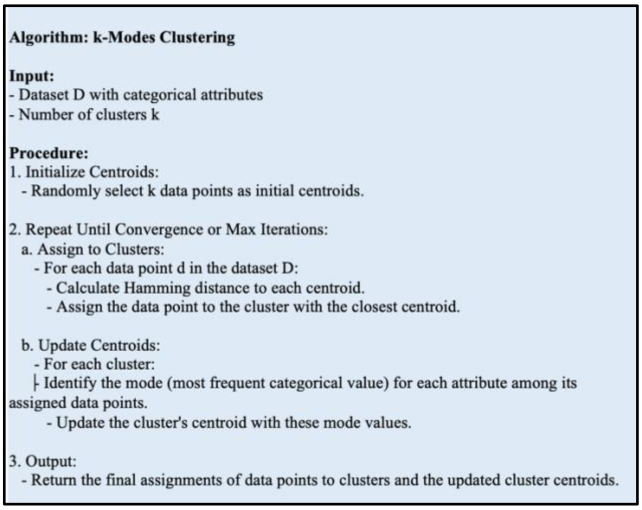
Figure 2: Cleaned BAPD Dataset

Furthermore, the Inscriptions attribute initially contained the actual textual inscription embedded on the pots and about 65% of the rows had Null value for this attribute. Since it is difficult to predict or impute actual inscriptions but at the same time the presence or absence of inscriptions is an important indicator, a modification was made to this column. A Boolean 1 was put to indicate the presence of inscription and a 0 was put to indicate the presence of inscriptions. Next, after performing preliminary analysis on the attributes, it was observed that all the selected columns had categorical values. The Fabric column had very skewed values with about 3/4th of them having the value “Athenian”. Further investigation into the dataset revealed that since the data is primarily about Greek pottery most of which was created using clay from within the region, a majority of artefacts having the Fabric value of “Athenian” makes sense. However, since this might not contribute to trend identification since the value is pretty uniform , this column was ultimately not considered in the classification task.

Further, to prevent multiple representations of the same item, the fuzzywuzzy library in python was used for approximate string matching. This library uses the Levenshtein distance algorithm in tandem with a threshold value to calculate the minimum number of single character edits that would be required to change one string to another. This string matching helps to group together similar categorical values to have cohesive non repetitive categories in the dataset. This resulted on the final dataset that was used for K-modes clustering.

3.1.1. The KModes Algorithm and Cluster Generation

For the present application, the K-modes clustering algorithm was selected. This was because the entire dataset is composed of categorical values and K-modes is an algorithm specifically designed to handle this type of data. Unlike traditional clustering algorithms that use distance metrics, K-modes uses the most frequent values within each cluster to determine the centroid. A sample pseudocode outlining the algorithm is shown in the figure below.



```
Algorithm: k-Modes Clustering

Input:
- Dataset D with categorical attributes
- Number of clusters k

Procedure:
1. Initialize Centroids:
  - Randomly select k data points as initial centroids.

2. Repeat Until Convergence or Max Iterations:
  a. Assign to Clusters:
    - For each data point d in the dataset D:
      - Calculate Hamming distance to each centroid.
      - Assign the data point to the cluster with the closest centroid.

  b. Update Centroids:
    - For each cluster:
    ├ Identify the mode (most frequent categorical value) for each attribute among its
assigned data points.
      - Update the cluster's centroid with these mode values.

3. Output:
  - Return the final assignments of data points to clusters and the updated cluster centroids.
```

Figure 3: Pseudocode For The K-Modes Clustering Algorithm

Just like all k parameter based clustering algorithms, the choice if k is a pivitol consideration. For the present application, we divide the dataset into sections based time periods and then ran a silhouette score analysis on the above chunks of the data. The results produced by the Silhouette analysis are shown in the figure below.

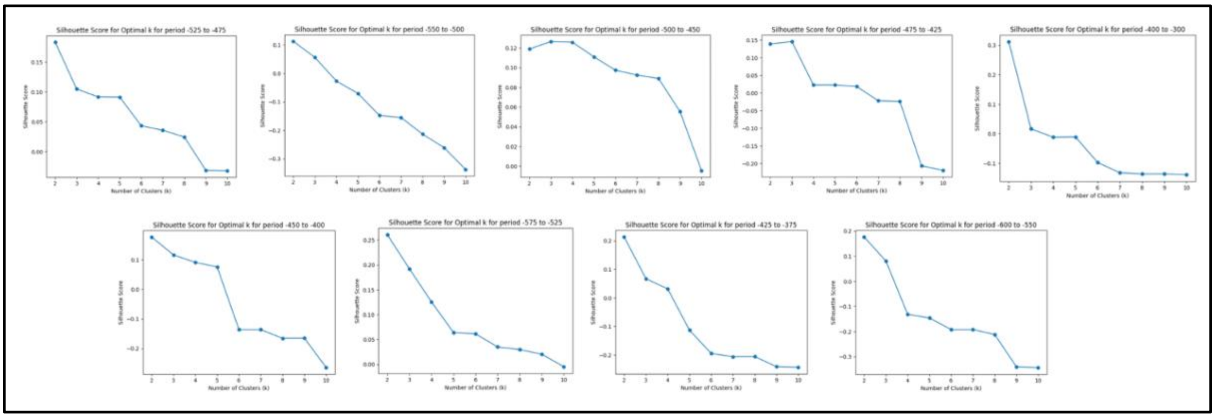
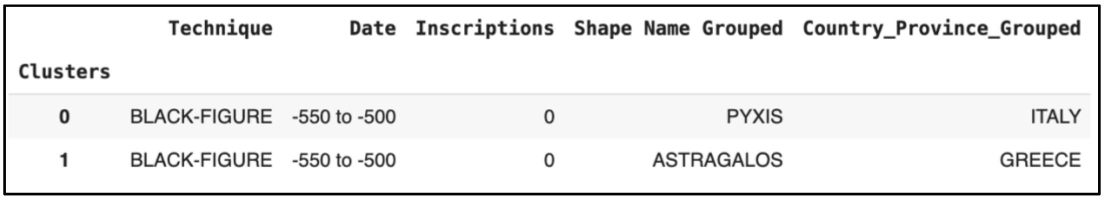


Figure 4: Silhouette Score For Different Time Based Subsets Of The BAPD Dataset

Observing the silhouette scores and consequently the number of optimal clusters, it can be observed that while for most of the defined time period, two clusters remain the standard distribution, there is a period of 4 time frames where there is a slight shift observed and we thus focus our attention to these periods. From -550 to -500 the optimal number of clusters is 2 but then from -500 to -450, the number of optimal clusters jumps to 4. From -475 to -425 this value falls to 3 and then again to 2 in -400 to -300. The optimal cluster number remains pretty uniform at 2 for the next time periods. Based on the obtained optimal K values, the KModes clustering analysis was performed separately for all these k periods and the most common behaviour for each of the clusters in each of the time periods was identified. A detailed presentation of the results and their implications are discussed in the section below.

3.1.2. Results and Implications

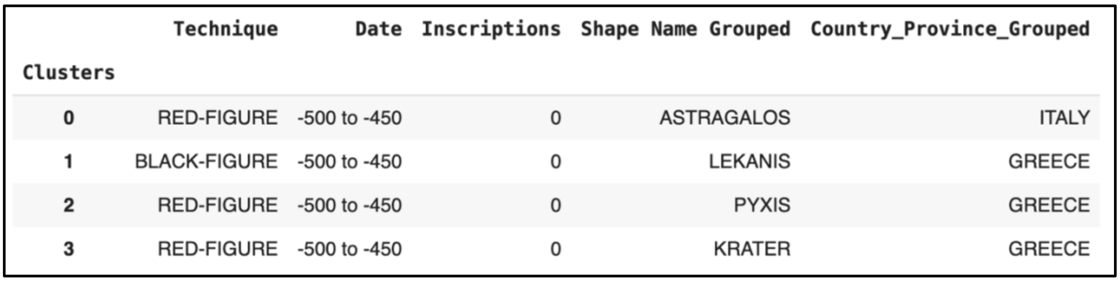
First, for the period from -550 to -500 the number of observed optimal k value is 2 and after clustering, the centroids for the 2 clusters observed are shown in the image below.



	Technique	Date	Inscriptions	Shape Name Grouped	Country_Province_Grouped
Clusters					
0	BLACK-FIGURE	-550 to -500	0	PYXIS	ITALY
1	BLACK-FIGURE	-550 to -500	0	ASTRAGALOS	GREECE

Figure 5: Centroids For K-Modes Clusters For The Time Period -550 To -500

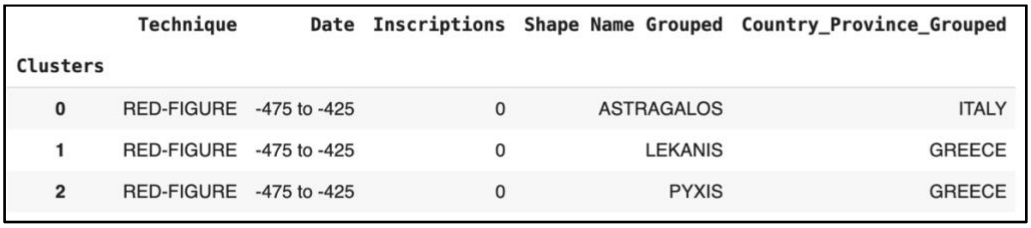
For the period from -500 to -450, the number of optimal clusters jumps to 4 and after applying K-modes clustering, the centroids of the clusters are shown in the image below.



	Technique	Date	Inscriptions	Shape Name Grouped	Country_Province_Grouped
Clusters					
0	RED-FIGURE	-500 to -450	0	ASTRAGALOS	ITALY
1	BLACK-FIGURE	-500 to -450	0	LEKANIS	GREECE
2	RED-FIGURE	-500 to -450	0	PYXIS	GREECE
3	RED-FIGURE	-500 to -450	0	KRATER	GREECE

Figure 6: Centroids For K-Modes Clusters For The Time Period -500 To -450

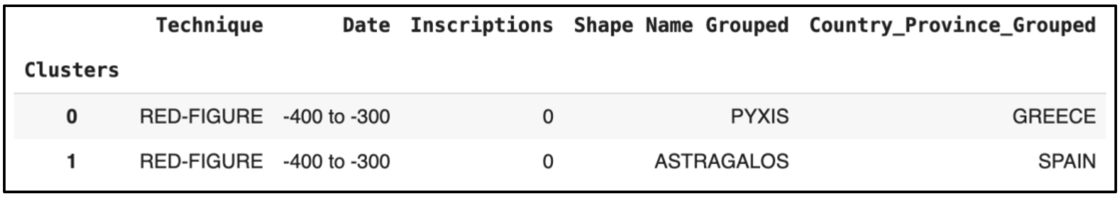
From -475 to -425 this value falls to 3 and after applying K-modes clustering, the centroids of the clusters are shown in the image below.



Clusters	Technique	Date	Inscriptions	Shape Name Grouped	Country_Province_Grouped
0	RED-FIGURE	-475 to -425	0	ASTRAGALOS	ITALY
1	RED-FIGURE	-475 to -425	0	LEKANIS	GREECE
2	RED-FIGURE	-475 to -425	0	PYXIS	GREECE

Figure 7: Centroids For K-Modes Clusters For The Time Period -475 To -425

In -400 to -300 it falls down back to 2 and the centroid of the clusters are shown in the image below.



Clusters	Technique	Date	Inscriptions	Shape Name Grouped	Country_Province_Grouped
0	RED-FIGURE	-400 to -300	0	PYXIS	GREECE
1	RED-FIGURE	-400 to -300	0	ASTRAGALOS	SPAIN

Figure 8: Centroids For K-Modes Clusters For The Time Period -400 To -300

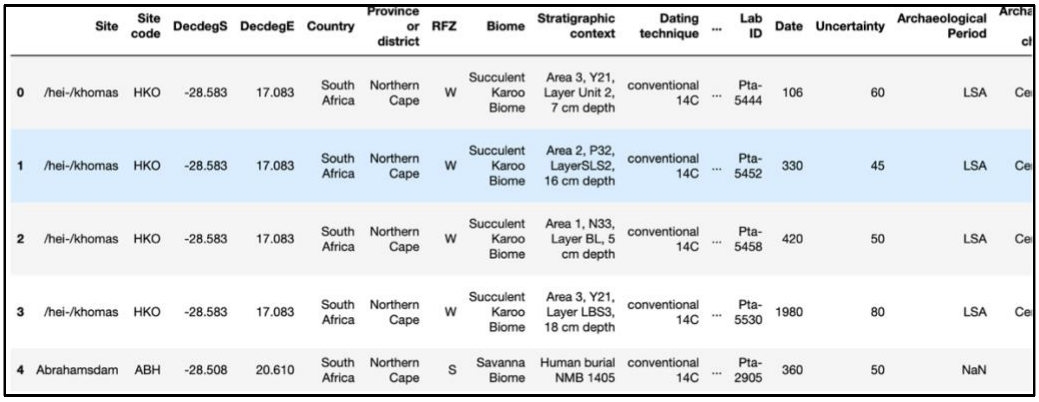
From the results of the above clustering exercise, it can be seen that in the period from -550 to -500, there are two main groups of Greek vases identified. The first is Pyxis shaped vases from Italy with no inscriptions and black figure painting and the second is Astragalos shaped vases from Greece with no inscriptions and black painted figures suggesting similar styles but with distinct characteristic elements. This is indicative of cultural and goods exchange between the two major identified communities – Archaic period Greece and Italy. Thus observation is also backed by the existence of external evidence suggesting that while these were distinct communities, significant trading was underway between them.

The clustering results of the period from -500 to -450 shows a sudden jump in the identified pottery clusters. It indicates now that in Italy, Astragalos shaped pottery with red painted figures and no inscriptions was prevalent - a shape and design characteristic of Greece in the previous year frame. Within Greece itself emergence of new shapes and designs can be seen indicating diversification of pottery production techniques. The next time period from -475 to -425 is indicative of a decline in Krater shaped pottery which is an ewer like vessel used for mixing vine with water. This could be indicative of a shift in social practices and preferences. Lastly, from -400 to -300 BCE, the emergence of Red-figure Astragalos from Spain and Red-figure Pyxis from Greece shows the further diversification in pottery styles possibly indicating the dynamic nature of artistic expression and cultural diffusion across different regions within this time period.

Thus, in conclusion it can be said that, the clustering analysis of ancient Greek pottery data reveals distinct patterns in evolution, indicating cultural exchanges between Greece and Italy, shifts in artistic preferences, and diversification of production techniques over various historical periods. The clustering results offer valuable insights into the dynamic nature of ancient societies and their interconnected artistic traditions reinforcing the utility of data mining techniques in the field of archaeology for pattern detection and hypothesis support.

3.2. Case Study 2 – Using Classification to Date South African Artifacts

This case study utilizes the Southern African Radio Carbon dataset as maintained by the Radiocarbon Accelerator Unit at the University of Oxford along with a random forest classifier in order to predict the Archeologic Period to which an artefact belongs. The South African Radio Carbon Dataset or SARD is an open-access online data repository maintained at the University of Oxford's Radio Carbon Accelerator unit or ORAU [55]. The dataset contains information on materials and their radiocarbon dates from South African Archaeological sites. The raw form of the data is illustrated in the figure below. Initially the dataset constituted of 22 attributes. This is shown in the figure below.



	Site	Site code	DecdegS	DecdegE	Country	Province or district	RFZ	Biome	Stratigraphic context	Dating technique	...	Lab ID	Date	Uncertainty	Archaeological Period	Archa ch
0	/hei-/khomas	HKO	-28.583	17.083	South Africa	Northern Cape	W	Succulent Karoo Biome	Area 3, Y21, Layer Unit 2, 7 cm depth	conventional 14C	...	Pta-5444	106	60	LSA	Ce
1	/hei-/khomas	HKO	-28.583	17.083	South Africa	Northern Cape	W	Succulent Karoo Biome	Area 2, P32, LayerSLS2, 16 cm depth	conventional 14C	...	Pta-5452	330	45	LSA	Ce
2	/hei-/khomas	HKO	-28.583	17.083	South Africa	Northern Cape	W	Succulent Karoo Biome	Area 1, N33, Layer BL, 5 cm depth	conventional 14C	...	Pta-5458	420	50	LSA	Ce
3	/hei-/khomas	HKO	-28.583	17.083	South Africa	Northern Cape	W	Succulent Karoo Biome	Area 3, Y21, Layer LBS3, 18 cm depth	conventional 14C	...	Pta-5530	1980	80	LSA	Ce
4	Abrahamsdam	ABH	-28.508	20.610	South Africa	Northern Cape	S	Savanna Biome	Human burial NMB 1405	conventional 14C	...	Pta-2905	360	50	NaN	

Figure 9: An Initial Snapshot Of The SARD Dataset

For the present case study only a subset of these attributes was considered and the retained attributes list is ['Country','Province or district', 'Biome', 'Dating technique', 'Material dated', 'Archaeological Period', 'Site Type']. From within the selected columns further cleaning and pre-processing had to be performed to prepare the data for the classification random forest model. Firstly, he country and province attributes were merged to obtain a unique location identifier. Next, the fuzzywuzzy library was employed for the removal of redundant material values through fuzzy logic. Finally, the missing values in the Archaeological Period Attribute were taken care of using mode imputation. Following, the pre-processing the data distribution of the target variable - the archaeological period - was visualized to get a better gauge on what type of decision tree model would work best. This is shown in the figure below.

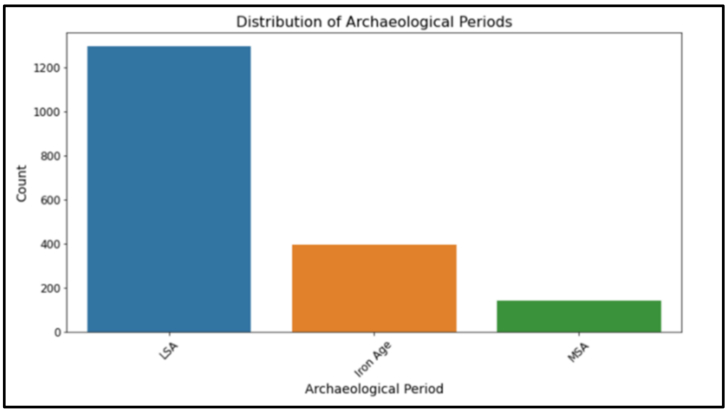
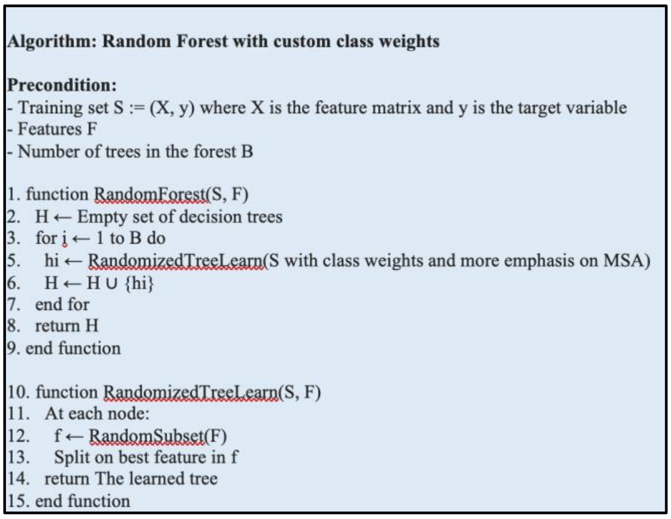


Figure 10: Distribution Of The Categorical "Archaeological Period" Attribute

As can be seen from the figure above, the categorical data is skewed in nature and this issue needs to be addressed to build a sound classifier. The method used to address of this issue, as well as the decision tree model that was built are described in detail in the section that follows.

3.2.1. Addressing Skewness and building the decision tree classifier

Random forest classifiers are a class of ensemble learning classification algorithms that work on the concept of using multiple weak learners and combining their classification power to create a strong and sound prediction model. In our case, a random forest model with 100 decision trees was created and bagging or bootstrap aggregation was used in order to combine the results of the prediction to produce the final classification results. Furthermore, since the data in our case is skewed, we use class weights to improve the performance of the low frequency classes. The pseudocode for the mechanism used for the creation of the present random forest classifier is provided below.



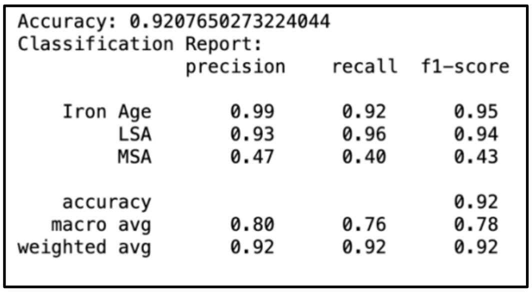
```
Algorithm: Random Forest with custom class weights

Precondition:
- Training set S := (X, y) where X is the feature matrix and y is the target variable
- Features F
- Number of trees in the forest B

function RandomForest(S, F)
  H ← Empty set of decision trees
  for i ← 1 to B do
    hi ← RandomizedTreeLearn(S with class weights and more emphasis on MSA)
    H ← H ∪ {hi}
  end for
  return H
end function

function RandomizedTreeLearn(S, F)
  At each node:
    f ← RandomSubset(F)
    Split on best feature in f
  return The learned tree
end function
```

Figure 11: Pseudocode For The Random Forest Classifier

The class weights parameter is used while creating the random forest classifier. While the two biggest classes in the target variable are given the same weight, the class with the least number of samples is given a higher weight. The results of the classification using the random forest thus created are presented in the section that follows.

3.2.2. Classifier results and metrics

The random forest classifier created above was trained on test and train subsections of the data. A ratio of 80-20 was used for the test train split exercise. The model was suitably trained and the results of the classification are presented below.



```
Accuracy: 0.9207650273224044
Classification Report:
              precision    recall  f1-score

    Iron Age       0.99      0.92      0.95
         LSA       0.93      0.96      0.94
         MSA       0.47      0.40      0.43

    accuracy                           0.92
   macro avg       0.80      0.76      0.78
weighted avg       0.92      0.92      0.92
```

Figure 12: Classification Results From The Random Forest Model

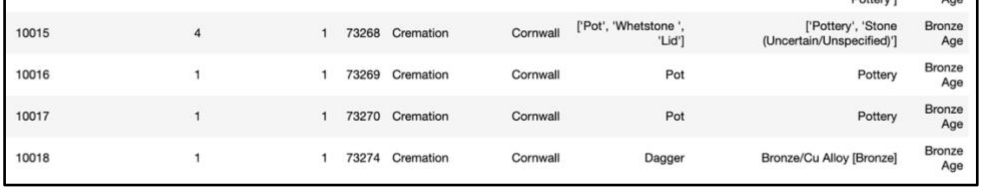
As can be seen from the figure above, the overall accuracy of the model is approximately 92%. This marks an overall 10% increase in accuracy from the case where oversampling was used to address the target attribute's class imbalance indicating that for our present application, the class eight approach of handling skewness works better. As can be seen from the results presented in the figure above, the precision which is a measure of the accuracy of positive predictions, about 99% of the instances were predicted as positive correctly whereas for LSA this number was 93% of the samples. In case of MSA however, this value is much lower at 47%.

Similarly for the Recall that evaluates the proportion of true positive instances correctly identified by the model, The performance of Iron Age and LSA class is good at 92% and 96% respectively whereas the performance of the model for the MSA class is quite low at 40%. The F1 score which indicates the proportion of correct predictions made across the model values are similarly high for Iron Age and LSA. This indicates that despite oversampling and class weight metrics, the imbalance in class while not significantly impacting the overall model performance, remains a significant disadvantage of our model. And while these skewness handling techniques do help in improving the prediction capabilities in, the 89% difference in values between MSA and the Iron age class in terms of samples can only truly be bridged using more data. Overall, using the SARD we successfully built a reasonably performing model that given characteristics about a sample can act as a precursor to complicated radio carbon dating procedures and offer initial indicators about the time period of origin of an unearthed artifact.

3.3. Case Study 3 – Association Rule Mining on Buried Grave Goods

In this case study, the data repository maintained at the Digital Archive for Grave Goods: Objects and Death in Later Prehistoric Britain as maintained by the UK's Archaeology Data Service was used. This database contains information on grave artifacts from the Neolithic, Bronze Age, and Iron Age [56]. The objective of the study is to use the dataset to explore relationships amongst the burial site attributes and the grave goods using association rule mining to identify the significance of specific goods occurrences within distinct cultural contexts and periods resulting in a holistic view of burial practices in older civilizations.

The dataset contains information on the goods found in excavated graves from the Neolithic, Bronze Age, and Iron Age Britain, roughly from 4000 BC to AD 43. The database itself is made up of multiple distinct csv files that can be linked together to create the entire comprehensive data set. At first, each of these files was pre-processed and cleaned. Fuzzy string matching was used to standardise the categorical values and mode imputation was used to handle missing values. Following this, python programming language and the joins and group by capability offered by its pandas library were used to combine the distinct csv files into one central data frame. This final dataset with relevant attributes is illustrated in the figure below.



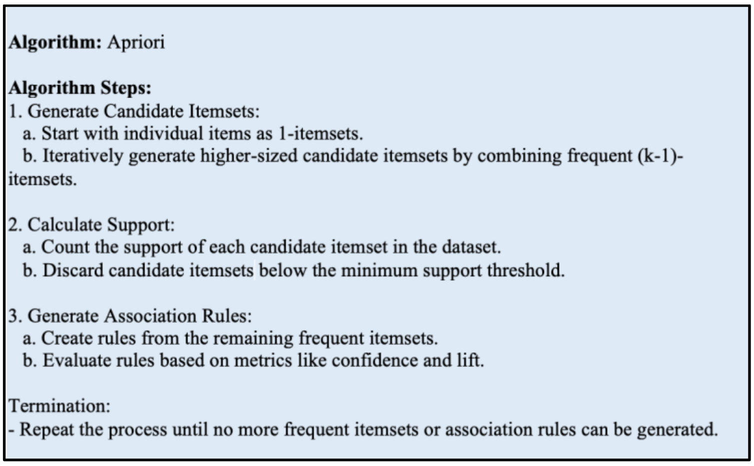
GID	Number_object_records	Number_hr_records	GID_hr	HR_type	Case_study_area	Object_type	Materials_summary	Period
10013	2	1	73266	Cremation	Cornwall	['Pot', 'Basket']	['Pottery', 'Organic (Uncertain/Unspecified)']	Bronze Age
10014	2	1	73267	Cremation	Cornwall	['Pot', 'Knife']	['Bronze/Cu Alloy [Bronze]', 'Pottery']	Bronze Age
10015	4	1	73268	Cremation	Cornwall	['Pot', 'Whetstone ', 'Lid']	['Pottery', 'Stone (Uncertain/Unspecified)']	Bronze Age
10016	1	1	73269	Cremation	Cornwall	Pot	Pottery	Bronze Age
10017	1	1	73270	Cremation	Cornwall	Pot	Pottery	Bronze Age
10018	1	1	73274	Cremation	Cornwall	Dagger	Bronze/Cu Alloy [Bronze]	Bronze Age

Figure 13: Constructed Grave Goods Dataset

The dataset had information on graves distributed across the following regions in the UK - Cornwall and the Isles of Scilly, Dorset, Kent, East Yorkshire, Gwynedd and Anglesey, Orkney and the Outer Hebrides. From within these regions, four geographically distinct areas namely Cornwall, Kent, Orkney and the Outer Hebrides, and East Yorkshire were selected for frequent item set mining. The data for each of these locations based on the 'Case_study_area' attribute was extracted separately and Apriori algorithm was run on this dataset to identify frequent goods sets buried with human remains across the study sites. The application of the algorithm along with the results are discussed in the section that follows.

3.3.1. Application of Apriori algorithm to mine frequent object sets

The Apriori algorithm is one of the central techniques in data mining frequent item set identification. Given a list of transaction like entries, this algorithm can effectively help recognise items that most frequently occur together across the entirety of the provided data. The key idea of Apriori is the "apriori property," which states that if an itemset is frequent, then all of its subsets must also be frequent. This property allows the algorithm to prune the search space, making it more efficient. A Pseudocode giving an overview of the working of the Apriori algorithm is provided below.



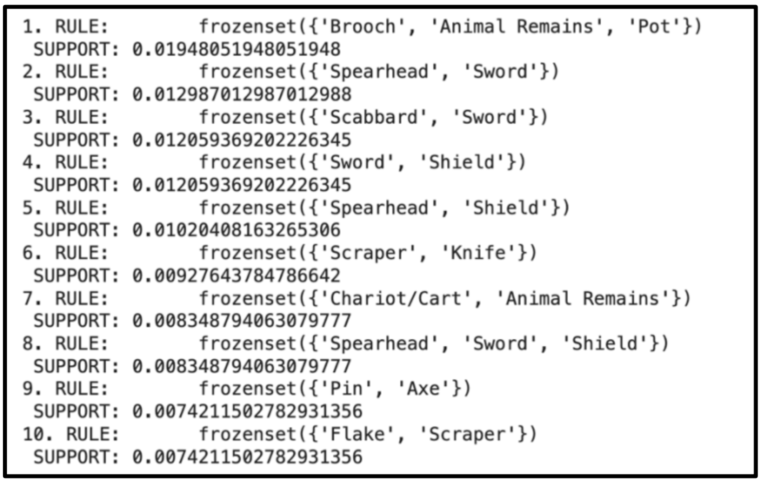
```
Algorithm: Apriori

Algorithm Steps:
1. Generate Candidate Itemsets:
   a. Start with individual items as 1-itemsets.
   b. Iteratively generate higher-sized candidate itemsets by combining frequent (k-1)-
itemsets.

2. Calculate Support:
   a. Count the support of each candidate itemset in the dataset.
   b. Discard candidate itemsets below the minimum support threshold.

3. Generate Association Rules:
   a. Create rules from the remaining frequent itemsets.
   b. Evaluate rules based on metrics like confidence and lift.

Termination:
- Repeat the process until no more frequent itemsets or association rules can be generated.
```

Figure 14: Pseudocode For The Apriori Algorithm

In the present case study, we try to draw on this "apriori property" quality to identify frequently occurring grave goods and study the implications of the results in the geographical context. The apyori package's Apriori library in python programming language was used in order to implement this data on the given dataset. The results produced for the selected four geographical regions and the implication of the results is discussed in the next section.

3.3.2. Results and Implications

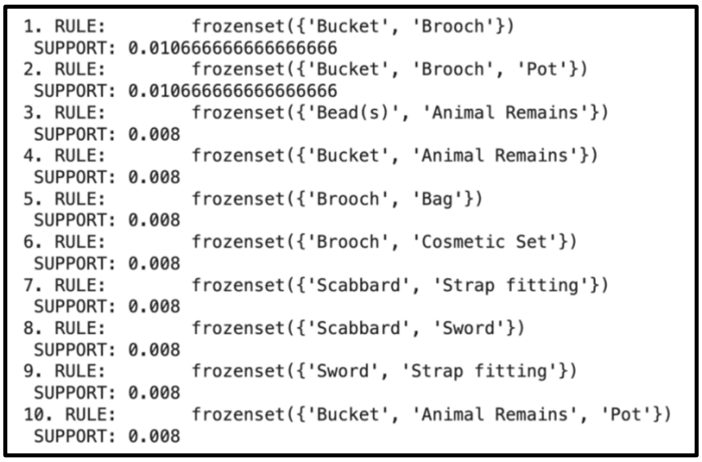
Starting with data from the region of East Yorkshire, the application of the Apriori algorithm produced the frequent item sets shown in the figure below along with associated support values.



```
1. RULE:        frozenset({'Brooch', 'Animal Remains', 'Pot'})
 SUPPORT: 0.01948051948051948
2. RULE:        frozenset({'Spearhead', 'Sword'})
 SUPPORT: 0.012987012987012988
3. RULE:        frozenset({'Scabbard', 'Sword'})
 SUPPORT: 0.012059369202226345
4. RULE:        frozenset({'Sword', 'Shield'})
 SUPPORT: 0.012059369202226345
5. RULE:        frozenset({'Spearhead', 'Shield'})
 SUPPORT: 0.01020408163265306
6. RULE:        frozenset({'Scraper', 'Knife'})
 SUPPORT: 0.00927643784786642
7. RULE:        frozenset({'Chariot/Cart', 'Animal Remains'})
 SUPPORT: 0.008348794063079777
8. RULE:        frozenset({'Spearhead', 'Sword', 'Shield'})
 SUPPORT: 0.008348794063079777
9. RULE:        frozenset({'Pin', 'Axe'})
 SUPPORT: 0.007421150278293135 6
10. RULE:       frozenset({'Flake', 'Scraper'})
 SUPPORT: 0.0074211502782931356
```

Figure 15: Frequent Item Sets For The Region Of East Yorkshire

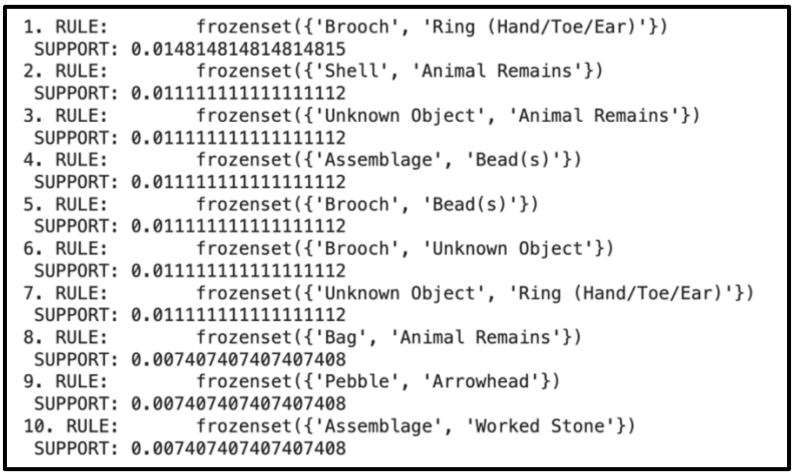
In this case, the presence of the set (Brooch, Animal remains, Pot) might suggest a burial practice where individuals were adorned with brooches, and animal remains and pots were included in the burial as offerings or for symbolic purposes. Further, he presence of spearheads and swords together could indicate a warrior burial or a community with a strong martial tradition. The presence of (Scabbard, Sword) set reinforces the martial aspect, suggesting the inclusion of sword-related items in burials. In addition the presence of (Scraper, Knife) set might suggest a burial related to craftsmanship or daily activities involving tools. The presence of a chariot or cart along with animal remains might indicate a burial with elements of transportation and perhaps a higher social status. Next, for the region of Kent, the item sets shown in the figure below were identified.



```
1. RULE:        frozenset({'Bucket', 'Brooch'})
 SUPPORT: 0.010666666666666666
2. RULE:        frozenset({'Bucket', 'Brooch', 'Pot'})
 SUPPORT: 0.010666666666666666
3. RULE:        frozenset({'Bead(s)', 'Animal Remains'})
 SUPPORT: 0.008
4. RULE:        frozenset({'Bucket', 'Animal Remains'})
 SUPPORT: 0.008
5. RULE:        frozenset({'Brooch', 'Bag'})
 SUPPORT: 0.008
6. RULE:        frozenset({'Brooch', 'Cosmetic Set'})
 SUPPORT: 0.008
7. RULE:        frozenset({'Scabbard', 'Strap fitting'})
 SUPPORT: 0.008
8. RULE:        frozenset({'Scabbard', 'Sword'})
 SUPPORT: 0.008
9. RULE:        frozenset({'Sword', 'Strap fitting'})
 SUPPORT: 0.008
10. RULE:       frozenset({'Bucket', 'Animal Remains', 'Pot'})
 SUPPORT: 0.008
```

Figure 16: Frequent Item Sets For The Region Of Kent

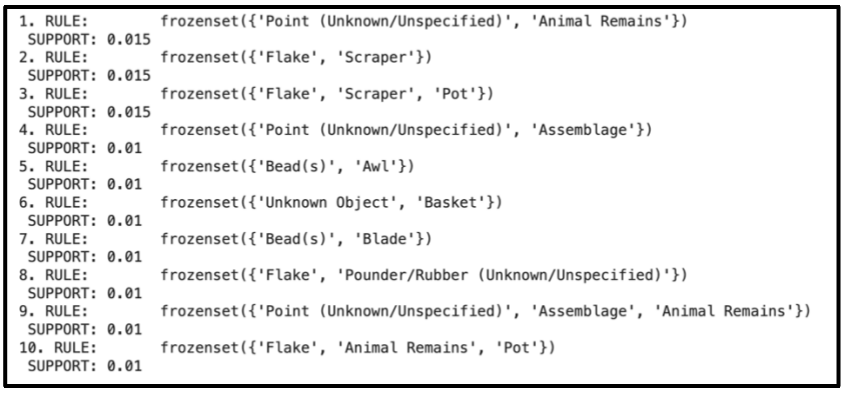
In the case of item sets identified for ancient graves in Kent, the presence of the combination of a bucket and brooch might suggest a burial with items associated with personal adornment and daily use. The identification of the (Beads, Animal remains) set could further indicate a burial with a focus on personal ornamentation and the importance of animals in the cultural or economic context. The combination of (Bucket, Animal remains, Pot) suggests a burial with a mix of functional and symbolic artifacts. For the region of Cornwall, the frequent sets shown in the figure below were identified.



```
1. RULE:        frozenset({'Brooch', 'Ring (Hand/Toe/Ear)'})
 SUPPORT: 0.014814814814814815
2. RULE:        frozenset({'Shell', 'Animal Remains'})
 SUPPORT: 0.011111111111111112
3. RULE:        frozenset({'Unknown Object', 'Animal Remains'})
 SUPPORT: 0.011111111111111112
4. RULE:        frozenset({'Assemblage', 'Bead(s)'})
 SUPPORT: 0.011111111111111112
5. RULE:        frozenset({'Brooch', 'Bead(s)'})
 SUPPORT: 0.011111111111111112
6. RULE:        frozenset({'Brooch', 'Unknown Object'})
 SUPPORT: 0.011111111111111112
7. RULE:        frozenset({'Unknown Object', 'Ring (Hand/Toe/Ear)'})
 SUPPORT: 0.011111111111111112
8. RULE:        frozenset({'Bag', 'Animal Remains'})
 SUPPORT: 0.007407407407407408
9. RULE:        frozenset({'Pebble', 'Arrowhead'})
 SUPPORT: 0.007407407407407408
10. RULE:       frozenset({'Assemblage', 'Worked Stone'})
 SUPPORT: 0.007407407407407408
```

Figure 17: Frequent Item Sets For The Region Of Cornwall

For Cornwall, the presence of (Brooch, Rings) set suggests a burial with items associated with personal adornment, indicating a focus on aesthetics – similar to Kent. The presence of shells and animal remains might signify a burial with items related to the natural environment or possibly ritual practices. The (Pebble, Arrowhead) set is intriguing and may suggest a burial with items associated with hunting or ritual significance. Finally for the region of Orkney and Outer Hebrides, the frequent item sets shown in the figure below were identified.



```
1. RULE:        frozenset({'Point (Unknown/Unspecified)', 'Animal Remains'})
 SUPPORT: 0.015
2. RULE:        frozenset({'Flake', 'Scraper'})
 SUPPORT: 0.015
3. RULE:        frozenset({'Flake', 'Scraper', 'Pot'})
 SUPPORT: 0.015
4. RULE:        frozenset({'Point (Unknown/Unspecified)', 'Assemblage'})
 SUPPORT: 0.01
5. RULE:        frozenset({'Bead(s)', 'Awl'})
 SUPPORT: 0.01
6. RULE:        frozenset({'Unknown Object', 'Basket'})
 SUPPORT: 0.01
7. RULE:        frozenset({'Bead(s)', 'Blade'})
 SUPPORT: 0.01
8. RULE:        frozenset({'Flake', 'Pounder/Rubber (Unknown/Unspecified)'})
 SUPPORT: 0.01
9. RULE:        frozenset({'Point (Unknown/Unspecified)', 'Assemblage', 'Animal Remains'})
 SUPPORT: 0.01
10. RULE:       frozenset({'Flake', 'Animal Remains', 'Pot'})
 SUPPORT: 0.01
```

Figure 18: Frequent Item Sets For The Region Of Orkney And Outer Hebrides

The item sets identified in this region are by far the most characteristic and definitive of the 4 sites studied. The (Point, Animal remains) set could indicate a burial associated with hunting or possibly ritual practices. The (Beads, Awl) and (Beads, Blade) sets suggest burials with items related to craftsmanship or personal ornamentation. The (Flake, Scraper) and (Flake, Scraper, Pot) sets suggest burials with items associated with tool use and possibly daily activities. Thus Orkney and Outer Hebrides burial finds suggest a population with a strong focus on hunting, craftsmanship, and daily tool use.

4. Conclusions

The amalgamation of ancient wisdom and contemporary analytics is reshaping archaeological research. With a myriad of data now accessible, mining and analysis techniques serve as a bridge between antiquity and the data-driven age, enriching our comprehension of human history and culture in unprecedented ways. In this article, we provided a detailed overview of utility of using data mining principles in archaeology. An overview of the historical relationship between the two fields along with the factors influencing research development was provided. Additionally, a detailed survey of the current status of research was conducted and a brief overview of some of the most promising work in the field was presented. Further, in order to demonstrate the advantage of the use of data mining algorithms in archaeology data, three comprehensive case studies were also presented each of which led to interested pattern observations.

The findings of this article underscore the collaborative nature of interdisciplinary research, where data mining becomes a tool to uncover societal structures and cultural exchanges throughout history. And while it is evident that the exploration of the past through analytical frameworks is an ongoing discourse, as data mining in archaeology continues to illuminate hidden facts from our collective heritage, the excitement of venturing into new territories through these tools is only increasing.

5. References

[1] Duarte, F. (2023, April 3). Amount of data created daily (2023). *Exploding Topics*. https://explodingtopics.com/blog/data-generated-per-day

[2] Schlanger, S. H., Wilshusen, R. H., & Roberts, H. (2015). From Mining Sites to Mining Data: Archaeology’s Future. The Kiva, 81(1–2), 80–99. https://doi.org/10.1080/00231940.2015.1118739

[3] Puyol-Gruart, Josep. (2002). Computer Science, Artificial Intelligence and Archaeology. *Archaeologists use lidar technology to map wealth and status in ancient Maya society*. (n.d.). Tulane News. https://news.tulane.edu/news/archaeologists-use-lidar-technology-map-wealth-and-status-ancient-maya-society

[4] Saraceni, J. E. (n.d.). *New 3-D facial reconstruction of Tutankhamun released - Archaeology Magazine*. https://www.archaeology.org/news/11492-230608-tutankhamun-facial-approximation

[5] *Spatial analysis in Archaeology*. (n.d.). Google Books. https://books.google.com/books/about/Spatial_Analysis_in_Archaeology.html?id=dgQ4AAAAIAAJ

[6] Kintigh, K. (n.d.-b). *America’s archaeology data keeps disappearing – even though the law says the government is supposed to preserve it*. The Conversation. https://theconversation.com/americas-archaeology-data-keeps-disappearing-even-though-the-law-says-the-government-is-supposed-to-preserve-it-104674

[7] McManamon, F., Kintigh, K., Ellison, L. A., & Brin, A. (2017). TDAR. Advances in Archaeological Practice, 5(3), 238–249. https://doi.org/10.1017/aap.2017.18

[8] Planning for the creation of digital data – Archaeology Data Service. (n.d.). https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/the-project-lifecycle/planning-for-the-creation-of-digital-data/

[9] The Digital Archaeological Record. (2018, October 6). The Digital Archaeological Record. https://www.tdar.org/

[10] Huggett, J. (2020). Is big digital data different? Towards a new archaeological paradigm. Journal of Field Archaeology, 45(sup1), S8–S17. https://doi.org/10.1080/00934690.2020.1713281

[11] Ouellette, J. (2021, January 6). Archaeology is going digital to harness the power of Big Data. Ars Technica. https://arstechnica.com/science/2021/01/archaeology-is-going-digital-to-harness-the-power-of-big-data/

[12] *Data mining based on an archaeological geoinformation system ArGIS*. (2016, November 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7818975

[13] 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). *Microchemical Journal*, *88*(2), 150–159. https://doi.org/10.1016/j.microc.2007.11.012

[14] Romero, L. L., Benito-Calvo, A., & Ríos-Garaizar, J. (2021). Defining and Characterising Clusters in Palaeolithic Sites: a Review of Methods and Constraints. *Journal of Archaeological Method and Theory*, *29*(1), 305–333. https://doi.org/10.1007/s10816-021-09524-8

[15] Caspari, G., & Jendryke, M. (2017). Archsphere – A cluster algorithm for archaeological applications. *Journal of Archaeological Science: Reports*. https://doi.org/10.1016/j.jasrep.2017.05.052

[16] Dixon, Boyd & Gosser, Dennis & Williams, Scott. (2008). Traditional Hawaiian men's houses and their socio-political context in Lualualei, Leeward west O'Ahu, Hawai'i. Journal of the Polynesian Society. 117. 267-295.

[17] López-García, P., & Argote, D. L. (2023b). Cluster analysis for the selection of potential discriminatory variables and the identification of subgroups in archaeometry. Journal of Archaeological Science: Reports, 49, 104022. https://doi.org/10.1016/j.jasrep.2023.104022

[18] NEOS Server: PSWarm. (n.d.). https://neos-server.org/neos/solvers/go:PSwarm/AMPL.html

[19] S. Bi, S. Xue, Y. Xu and A. Pei, "Spatial Data Mining in Settlement Archaeological Databases Based on Vector Features," 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Jinan, China, 2008, pp. 277-281, doi: 10.1109/FSKD.2008.490.

[20] Rasheed, N. A., & Nordin, J. (2020). Classification and reconstruction algorithms for the archaeological fragments. Journal of King Saud University - *Computer and Information Sciences*, *32*(8), 883–894. https://doi.org/10.1016/j.jksuci.2018.09.019

[21] *Classification of 3D archaeological objects using Multi-View Curvature Structure signatures*. (2019). IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/8576529

[22] Makridis, M., & Daras, P. (2012). Automatic classification of archaeological pottery sherds. *Journal on Computing and Cultural Heritage*, *5*(4), 1–21. https://doi.org/10.1145/2399180.2399183

[23] Q. Li-Ying and W. Ke-Gang, "Kernel fuzzy clustering based classification of Ancient-Ceramic fragments," 2010 2nd IEEE International Conference on Information Management and Engineering, Chengdu, China, 2010, pp. 348-350, doi: 10.1109/ICIME.2010.5477818.

[24] Qu, F., Hu, Y., Yang, Y., & Gu, X. (2011). Unsupervised kernel fuzzy clustering based on differential evolution algorithm in intelligent Materials system. In Advances *in intelligent and soft computing* (pp. 189–192). https://doi.org/10.1007/978-3-642-23756-0_31

[25] Wells, E. C. (2014). Archaeometry: Definition. In *Springer eBooks* (pp. 468–470). https://doi.org/10.1007/978-1-4419-0465-2_360

[26] Kowalski, B. R., Schatzki, T. F., & Stross, F. H. (1972). Classification of archaeological artifacts by applying pattern recognition to trace element data. *Analytical Chemistry*, *44*(13), 2176–2180. https://doi.org/10.1021/ac60321a002

[27] Musumarra, G., Stella, M., Matteini, M., & Rízzí, M. (1995b). Multiariate characterization, using the SIMCA method, of mortars from two frescoes in Chiaravalle Abbey. *Thermochimica Acta*, *269–270*, 797–807. https://doi.org/10.1016/0040-6031(95)02533-2

[28] Baxter, M. (2006b). A Review Of Supervised And Unsupervised Pattern Recognition In Archaeometry*. *Archaeometry*, *48*(4), 671–694. https://doi.org/10.1111/j.1475-4754.2006.00280.x

[29] García-Heras, M., Blackman, M. J., Fernández-Ruiz, R., & Bishop, R. L. (2001b). Assessing Ceramic Compositional Data: A Comparison of Total Reflection X-ray Fluorescence and Instrumental Neutron Activation Analysis On Late Iron Age Spanish Celtiberian Ceramics. *Archaeometry*, *43*(3), 325–347. https://doi.org/10.1111/1475-4754.00020

[30] Charalambous, E., Δικωμίτου-ηλιάδου, M., Milis, G., Mitsis, G. D., & Ηλιάδης, Δ. Γ. (2016). An experimental design for the classification of archaeological ceramic data from Cyprus, and the tracing of inter-class relationships. *Journal of Archaeological Science: Reports*, *7*, 465–471. https://doi.org/10.1016/j.jasrep.2015.08.010

[31] Kobyliński, Ł., & Walczak, K. (2007). Data Mining Approach to classification of Archaeological aerial photographs. In Springer eBooks (pp. 479–487). https://doi.org/10.1007/3-540-33521-8_52

[32] Hao, Y., Liu, Y., Wu, Z., Han, L., Chen, Y., Chen, G., Chu, L., Tang, S., Yu, Z., Chen, Z., & Lai, B. (2021). EdgeFlow: Achieving Practical Interactive Segmentation with Edge-Guided Flow. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2109.09406

[33] Wilcke, W., De Boer, V., De Kleijn, M., Van Harmelen, F., & Scholten, H. (2019). User-centric pattern mining on knowledge graphs: An archaeological case study. Journal of Web Semantics, 59, 100486. https://doi.org/10.1016/j.websem.2018.12.004

[34] Barati, M., Bai, Q., & Liu, Q. (2016b). SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data. In Lecture Notes in Computer Science (pp. 30–43). https://doi.org/10.1007/978-3-319-42911-3_3

[35] Brown, M., & Shackel, P. A. (2023). Text Mining Oral Histories in Historical Archaeology. International Journal of Historical Archaeology, 27(3), 865–881. https://doi.org/10.1007/s10761-022-00680-5

[36] Baxter, M. J. (2015). Exploratory Multivariate Analysis in Archaeology. Eliot Werner Publications. https://doi.org/10.2307/j.ctv2sx9gfb

[37] Beazley Archive Pottery Database (BAPD). Accessed October 26, 2023. https://www.beazley.ox.ac.uk/carc/pottery.

[38] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

[39] Wang, W., Tan, Y., Jiang, J., Lu, J., Shen, G., & Yu, R. (2004b). Clustering based on kernel density estimation: nearest local maximum searching algorithm. Chemometrics and Intelligent Laboratory Systems, 72(1), 1–8. https://doi.org/10.1016/j.chemolab.2004.02.006

[40] CS221. (n.d.). https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[41] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Knowledge Discovery and Data Mining*.

[42] Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very fast C4.5 Decision Tree algorithm. Applied Artificial Intelligence, 32(2), 119–137. https://doi.org/10.1080/08839514.2018.1447479

[43] Wikipedia contributors. (2023b, November 18). Euclidean distance. Wikipedia. https://en.wikipedia.org/wiki/Euclidean_distance

[44] Hassan, M. U. (2023, May 9). VGG16 – Convolutional Network for Classification and Detection. Neurohive / Neural Networks. https://neurohive.io/en/popular-networks/vgg16/

[45] What is the k-nearest neighbors algorithm? | IBM. (n.d.). https://www.ibm.com/topics/knn

[46] Wikipedia contributors. (2023b, November 4). Support vector machine. Wikipedia. https://en.wikipedia.org/wiki/Support_vector_machine

[47] Musicant, D. R. (n.d.). Smooth support Vector machine home page. David R. Musicant. https://research.cs.wisc.edu/dmi/svm/ssvm/

[48] Wikipedia contributors. (2023b, October 16). Gabor wavelet. Wikipedia. https://en.wikipedia.org/wiki/Gabor_wavelet

[49] Chen, Z., & De B Harrington, P. (2019). Automatic soft independent modeling for class analogies. Analytica Chimica Acta, 1090, 47–56. https://doi.org/10.1016/j.aca.2019.09.035

[50] Wikipedia contributors. (2023c, November 3). Linear discriminant analysis. Wikipedia. https://en.wikipedia.org/wiki/Linear_discriminant_analysis

[51] Brownlee, J. (2020, August 14). Learning vector quantization for machine learning. MachineLearningMastery.com. https://machinelearningmastery.com/learning-vector-quantization-for-machine-learning/

[52] fuzzywuzzy. (2020, February 13). PyPI. https://pypi.org/project/fuzzywuzzy/

[53] What is Bagging? | IBM. (n.d.). https://www.ibm.com/topics/bagging

[54] Wikipedia contributors. (2023a, September 5). Apriori algorithm. Wikipedia. https://en.wikipedia.org/wiki/Apriori_algorithm

[55] Loftus, E., Mitchell, P., & Ramsey, C. B. (2019). An archaeological radiocarbon database for southern Africa. Antiquity, 93(370), 870–885. https://doi.org/10.15184/aqy.2019.75

[56] Anwen Cooper, Duncan Garrow, Catriona Gibson, Melanie Giles, Neil Wilkin, 2020. (updated 2023) https://doi.org/10.5284/1052206.