Sex Estimation of Vindolanda Roman Shoes Using Modern Footprint Data

By Adith Natarajan

September 2023

Abstract

Thousands of shoes were excavated from the Vindolanda Fort, a Roman auxiliary for located in Northumberland, England. The aim is to determine the efficacy of data mining approaches in sex estimation. Data mining techniques like feature extraction, pattern recognition and classification algorithms are employed to analyze and extract meaningful patterns and relationships from the foot and shoe data. Unlike traditional approaches, which rely mostly on foot length to estimate sex, our explanatory data analysis highlighted the importance of other foot features like Breadth at Ball, Breath at heel and their indices indicating unique correlations based on sex. The study is motivated for two reasons: first, using the shoes as a unique proxy for sex estimation in the world of archaeology, and, second, an investigation into the sex distribution at Vindolanda Fort. The classification models that were constructed for this research are Gaussian Naïve Bayes and classification tree methods that include pruning, random forest and XGBoost. The predictive models were trained using footprint measurements from European and Ghanaian populations and then they were applied to the Roman shoe data to predict sex. They achieved an average of 74% accuracy when validated with the Archeologist's predictions. Our analysis did conclude with predicting more females on the shoes found in Vindolanda Fort suggesting a potential underestimation of women's roles or presence. However, this study does highlight the necessity for a more extensive shoe dataset to validate these findings and further investigate the sex distribution of this Roman Fort.

Table of Contents

1. Introduction

In forensic anthropology, the identification of human remains is still a griping process, often fraught with challenges because of the poor conditions and degraded state of the skeletal remains that were unearthed after hundreds of years (De Boer et al., 2020). The first key aspect during the identification process is to determine the sex from the remains as it provides compelling details/ evidence for laying the foundation for a biological profile. As anthropological research has been limited in Vindolanda Fort due to a lack of well-preserved skeletons, traditional methods of sex estimation through cranial and pelvic bone analysis can be insufficient (Buck, et al., 2019). This lack of information makes it difficult to paint a comprehensive picture of the population's demographics. Thus, shoe sizes and footprints become extremely momentous for research, especially when the remains are not available or if they are found in a very fragile state. This paper aims to investigate the feasibility of the sex of a Roman based on shoe dimensions using machine learning techniques and also to contribute to the broader understanding of Roman military life.

Currently, data mining techniques have established confidence in forensic science, leveraging large datasets and progressive algorithms to recognize patterns and correlations between predictor variables (Delgado et al., 2021). According to a study completed in 2023, deep learning algorithms can be used to estimate sex using a peripheral Quantitative computed tomography (pQCT) slice of the fourth lumbar vertebra (L4) (Oura et al., 2023). These advancements show that machine learning techniques, specifically deep learning, have the potential to significantly improve the accuracy and efficiency of traditional forensic anthropology methods (Oura et al., 2023). Knecht et al. in 2023 applied machine learning classifiers to long bone measurements to estimate sex, particularly in the absence or degradation of more dimorphic bones. It aimed to increase the accuracy of sex estimation methods by generating and comparing various classification models (Knecht et al., 2023)

However, it is important to acknowledge the use of contemporary foot sizes by Europeans and Ghanaians for determining Roman shoes' sex is however subject to certain limitations. Additionally, there are considerable historical and regional variations in human stature that can affect the study. During the Roman period, statistics suggest that the stature of the population declined in areas such as Roman Britain and Roman Gaul (Quade, Gowland, 2021). Furthermore, stature can also vary significantly within these regions depending on factors such as rural or urban settings. On average, soldiers in specialized contexts like a Roman fort might have been taller than the local population or slaves due to their physical robustness. When comparing modern foot dimensions with those of ancient Roman shoes, such variations can introduce a level of uncertainty. Considering these limitations, caution should be exercised when estimating sex using data from contemporary feet. Hence, future parts of this study will aim to estimate and compare the mean stature derived from

Roman shoe dimensions with the mean stature of modern populations. By leveraging the wealth of limited data available from excavated shoes and impressions taken from modern European people and Ghanaians, we can try to bridge the gap between shoes and the people who once lived in Vindolanda Fort.

1.1 Aims and Objectives

1. ***Sex estimation using modern footprints:*** To utilize data from modern European and Ghanaian footprints for training predictive models aimed at estimating the sex of the possible owner of the Roman shoes based on their dimensions.
2. ***Model Evaluation against Archaeological predictions:*** To evaluate and compare the performance of the predictive models against sex predictions previously made by archaeologists for the Roman Shoes in the dataset.
3. ***Demographic insights inside Vindolanda Fort:*** To provide possible insights into the sex distribution at Vindolanda Fort, based on the outcomes of our predictive models.

1.2 Paper Structure

In this paper, there are five main sections. It begins with an introduction that introduces the topic and then moves to a background section that reviews previous research on Roman shoes and their relevance in archaeological evaluations. We will also be looking at traditional methods for sex estimation and introduce the theories proposed by Carol van Driel–Murray and Elizabeth M. Greene. We will explore the importance of shoes as a proxy for sex estimation and demographic exploration in Vindolanda Fort. Followed by the Methodology section which details the methods used for integrating European and Ghanian footprints for training the model. A key component of the paper is the analysis, modeling, and results section where we will conduct explanatory data analysis and analyze the models applied to the data, such as Naïve Bayes and tree-based algorithms, and they are compared at the end of this section. Lastly, in the conclusion section, the findings and implications are discussed and interpreted. We will also be looking at the limitations of this study and the challenges we overcame in using shoes for predicting sex.

2. Background

This fort is a historically significant archaeological site that is located south of Hadrian's Wall in Northern England. The site offers remarkable preservation conditions, which is one of its most impressive features. The Anaerobic conditions have preserved organic materials like leather exceptionally well, allowing us to gain a unique and invaluable insight into Roman life. In Vindolanda Fort, over 6,000 leather shoes preserved in these conditions offer an unprecedented look at the fort's ancient demographics. There are several reasons such as waterlogged conditions, burial and accumulations, and chemical and mineral interactions behind the preservation of Roman shoes found in Vindolanda Fort. It becomes crucial to understand these conditions before proceeding with the analysis since it provides valuable context and insights that can increase the accuracy and reliability of the analysis. The decomposing of wood and stone buildings was sealed with thick layers of clay during the Roman occupational period, forming layers in which oxygen was excluded from the decomposition (Orr et al., 2021). Waterlogging on dense clay layers resulted in the formation of anaerobic layers which provide optimum preservation conditions (Orr et al., 2021). The preservation conditions can significantly influence their physical characteristics and it becomes important to understand these conditions as data miners, to analyze the observed features and to account for any biases and alterations that may be introduced (Garvey, 2018). With this knowledge, the attributes of shoes can be analyzed more accurately and sex estimation models can be made more reliable. Researchers can also determine the most relevant and reliable features for sex estimation by understanding these preservation conditions. In some cases, features may degrade more rapidly under preservation conditions, so it is important to consider this information when selecting the most informative features that can effectively differentiate between male and female shoes.

2.1 Sex estimation in bioarchaeology and forensic anthropology

Traditionally, skeleton features such as the pelvis and post-cranial skeleton have been used to estimate sex in forensic anthropology and bioarchaeology (Sex Estimation of the Human Skeleton, 2020). It is common to use both metric and non-metric methods for sex estimation. Statistical models are used in metric studies to differentiate between male and female skeletons (Krishnan et al., 2016). On the other hand, nonmetric methods emphasize morphological characteristics (Krishnan et al., 2016). Factors that would define the morphology of the human foot are genetics, lifestyle, and climatic factors (Tomassoni et al., 2014). It is carried out through geometric morphometric analysis, and focuses on quantifying the shape of the skeletal features, thus offering a more refined understanding of skeletal structure that minimizes the influence of size (Krishnan et al., 2016). Another morphological method for sex estimation involves the examination of parturition markers. It is thought that these skeletal signs are caused by a physical strain that occurs during pregnancy and

childbirth (Sex Estimation of the Human Skeleton, 2020). However, there has been much discussion regarding their accuracy in predicting sex and their link with pregnancy (Sex Estimation of the Human Skeleton, 2020). Scientists have divergent opinions regarding their significance and accuracy in estimating biological sex (Sex Estimation of the Human Skeleton, 2020).

In most situations, demographic inferences are made from cemetery populations, which suffer from the Osteological Paradox, highlighting the problem of using samples from the dead to predict the characteristics of once-living populations (Buck et al., 2019). Since individuals of varying ages have different mortality rates, skeletal samples do not necessarily represent the population's demographic structure (Buck et al., 2019). As an alternative, footwear provides a more accurate representation of a living population, making it a useful proxy for skeletal remains. The archaeological record provides useful information for the identification of subadult populations.

Our method extends the contemporary footprint measurements to estimate sex based on insole shoe measurements, in contrast to traditional bioarchaeology and forensic methods that rely on skeletal metrics and osteological markers. Due to the lack of skeletal remains in Vindolanda, the traditional methods that I just described cannot be used. Considering these constraints, we used Roman shoe data as a proxy for the human body, providing an alternative way to estimate sex. Also, the use of Roman shoes allows us to work with a different type of archaeological material, expanding the number of factors we can use to estimate sex.

2.2 Women’s presence in military forts

One overlooked aspect of Vindolanda, however, is women’s and children’s role within this fort. According to traditional narratives, Roman forts are known to be predominantly masculine domains dominated by soldiers and commanders. However, it has been stressed in recent research that understanding children's and women’s roles in the fort could redefine our understanding of the culture and life that existed during this Roman period (1st to 5th century AD).

Most of the epigraphic evidence, which focuses almost exclusively on men, reinforces this view. Researchers have noted that only ten per cent of Roman British inscriptions feature women and these inscriptions offer uneven information, ranging from names etches on pottery to elaborate biographies (Allason-Jones, 2012). Due to the inherent bias in epigraphic records towards men, particularly those in military roles, women and children are hardly recognized. Therefore, in their narratives, women and children are often marginalized, which creates significant challenges for reconstructing the social fabric in Roman fort and surrounding communities.

Allason-Jones (1995), explores the question of whether the presence of women can be identified among a group of artefacts. In particular, she highlights the presence of jet items, which are most

commonly found in the graves of women (Allason -Jones, 1995). Besides indicating the presence of women, these could also reveal their roles and beliefs. The importance of jet items for women appears to have been correlated with religion or culture (Allason- Jones, 1995). Another physical property that was often related to women is spindle whorls, and some evidence indicates that they were used both by high-status and low-status female spinners within the fort (Alberti, 2018). Similarly, Greene (2016), asserts a similar argument by drawing upon evidence such as writing tablets, and shoes to argue that women were integral parts of this community. Several writing tablets mentioning travel and social visits indicate that some of the women enjoyed a degree of freedom and engaged in social activities (Greene, 2016). Hence, shoes provide a rich reservoir of biological and social information (Greene, 2016). The vital role of military diplomas offer personal information about soldiers and their families. Several documents reveal that women and children were part of these communities from the beginning, proving they were more prominent than traditionally thought. The examination of 42 military diplomas revealed that family connections were part of the military community, although they were overlooked to emphasize the male, soldierly nature of these environments (Greene, 2015).

The shoes found at Vindolanda are diverse, from high-end slippers to more functional styles, indicating that the society was much more than a collection of male soldiers. There were Certain shoe styles, such as sandals with open toes or delicate decorations, that have been traditionally associated with women (Van Driel- Murray, 2001). Therefore, the size variations can potentially differentiate between men's and women's average foot dimensions. Examining the sex distribution within the Roman population would provide more insight into its social dynamics and demographics within Hadrian's Wall. Using quantitative data to support or challenge traditional narratives, we attempt to correlate shoe size with the sex of the individual, extending Greene's work.

The foundation of our research lies in these initial studies. We add a quantitative dimension to a discussion largely driven by qualitative assessments by developing a methodology based on modern footprints. Increasing evidence points to the active participation of women and children in military settlements, especially through artefacts and written documents (Greene, 2015). We understand these communities as more than military outposts based on the presence of women and children evident in tablets and other finds. The work of Greene (2016) emphasizes this by highlighting that even criticism of women's presence in forts wasn't enough to ban them. It would only further solidify their role in these societies if a statistical basis for their presence were derived from shoe sizes.

Hence, we can uncover more robust, quantitative data about the demographic makeup by analyzing the sex distribution of ancient Roman shoes found in Vindolanda. Our methodology allows us to analyze a dataset of shoes recovered from such sites, which can reinforce or challenge the prevailing notion that these were male-dominated forts. Women and children were not only occasionally residing in these forts, but their existence was essential to the community as a whole (Greene, 2013). In this

case, it would be helpful if we could predict sex based on foot measurements in Vindolanda and possibly other forts of a similar type, so that we could add considerable weight to the argument that women and children were integrated into the culture of Vindolanda.

2.3 Using Shoes for determining sex

The identification and analysis of physical evidence is often the primary focus of forensic anthropology. Despite this, footprints and shoe evidence are becoming valuable yet underutilized sources of information. In a recent study on the modern population, they used footprint dimensions and unique morphological characteristics to determine sex and identify individuals (Awais et al., 2018). In the field of bioarchaeology, these techniques are applied largely to medico-legal and criminal investigations (Awais et al., 2018). This study from 2018, particularly highlights how footprint ridge density and other characteristics can be used to differentiate sexes, helping forensic investigators focus on areas of interest. Another study in 2007, demonstrates using various measurements of foot bones for estimating sex (Case and Ross, 2007). In this context, these forensic techniques appear to have intriguing potential applications in bioarchaeology, particularly in the study of Roman shoes. Lastly, a study from 2010 recorded the foot measurements and shoe sizes to determine the sex of an individual using statistical methods like univariate and multivariate discriminant function models (Atamturk, 2010). It yielded positive accuracy (82% -96%) rates and shoe length was identified as the most important variable (Atamturk, 2010).

Like modern footprints, Roman shoes can provide valuable anatomical details about the bodies they once represented. To predict the sex associated with these shoes, machine learning models can be trained using the foot data, similar to the studies mentioned above. These studies lay a foundation for our initiative to create machine-learning models tailored to bioarcheological applications. Once the model is constructed, trained, and validated against modern footprint data, it can be applied to the shoe data to predict the sex of the possible owner.

Therefore, the foot dimensions can serve as a proxy for the body, especially in our case where there is not much skeleton evidence to get a good demographic picture of the population. Note that there is only a small number of research that attempts to estimate sex by analysing footprint measurements, and even smaller when it comes to estimating sex using shoe dimensions. In this regard, advanced machine learning algorithms could make a significant contribution. Thus, as mentioned before in the absence of well-preserved skeletal remains, we depend on these shoes as a proxy which did not decompose due to the anaerobic layer.

2.4 Current approaches in assigning sex using foot dimensions

In a basic sex-assigning analysis for footwear, the key factor that is taken into account is the dimensions of the footwear. The comparison between modern foot size data and these shoes can be a great indicator of sex. However, from a statistical perspective, it is critical to consider that foot sizes and proportions can change exclusively, and it becomes challenging to decide the sexual orientation based on size. It is problematic due to the considerable overlap in foot sizes between men and women and it does not provide sufficient discriminatory control to differentiate the sex with a high level of confidence. Typically, the average male tends to have a larger foot compared to females which is influenced by several factors such as genetics, hormones, and skeletal bone patterns. During the development stages in childhood, boys' and girls' foot sizes start to differ around 10, and girls reach their full adult size between the ages of 11 and 13 while boys continue to grow until they reach the age of 15 -16 (Van – Driel Murray, 1995). The longer growth period results in an average difference of about 2cm in foot size between males and females (Van -Driel- Murray, 1995).

According to modern statistics, the course of growth has remained unchanged, but children are more advanced now than they were until the 19th century, suggesting that the divergence between boys and girls in antiquity may have started at a slightly later age, and boys may not have all reached adulthood until about 16 or 17 years of age (Van- Driel Murray, 1995). It has been suggested that the footwear excavated corresponded to modern European sizes ranging from 27 to 34, and were more likely worn by women (Van-Driel Murray, 1995). The women's foot sizes are likely to be at the lower end and there may have been men and women with sizes smaller than or equal to 34 (Van-Driel Murray, 1995). While these differences are present on average. Both men and women have a range of natural variations. These variations in foot anatomies do not necessarily reflect functional advantages or disadvantages. The morphological and metric differences are small between male and female skeletons and can be key aspects used for sex estimation purposes (Nikita and Nikitas, 2019). In a study that involved the factors affecting the medial longitudinal arch height of the foot in healthy young adults, the arch height where females typically exhibit a higher arc and a more pronounced curve along the inner side of the foot, and toe proportions where males tend to have wider toes (Nagano et al., 2018). There may be variations in balance, and overall foot functions caused by different toe proportions. Hallux Valgus angle is a condition that is the deviation of the big toe towards the other toes which also contributes to the structural differences between males and females. This occurs more commonly in females, and several factors have been reported to be associated with hallux valgus, including genetics, sex, age, BMI, foot pain, and the type of shoes they wear (Nagano et al., 2018). However, the ratio of arch height in females was positively correlated with the hallux valgus angle in this study (Nagano et al., 2018). Despite the value of basic foot measurements for

initial assessments, data mining techniques offer a more comprehensive, accurate, and robust method for estimating Roman shoe sex. With the help of data analysis, pattern recognition, and statistical methods, we can understand the complex relationships between various shoe characteristics and their associations with sex.

Van Driel- Murray was the first to investigate the demographics of the people in Roman forts using footwear as a metric to determine age and sex. She argued that Vindoalanda was more diverse which included women and children, by using shoe sizes and style rather than typical archaeological narratives that emphasized only the presence of Roman soldiers. She underlined the stereotypes about the population in the fort being connected with weapons and other military artefacts. On the other hand, Greene incorporated leather conservation techniques with Van Driel's ideas to study footwear assemblages taking size change into account. Greene also acknowledged the drawbacks of using shoes as the main indicator of demographics at the Roman Fort but was still able to challenge the conventional belief that the fort was male-dominated. Therefore, she reinforces the idea that Roman forts were diverse communities of men, women, and children by adding depth and methodological rigor to Van Driel's initial research.

2.5 Artificial Intelligence and Machine Learning for sex estimation

As machine learning algorithms have evolved over the last few years, they have become powerful tools for automation and improving classification problems. Archaeologists have used data mining techniques to focus on processing numerical and categorical data, textual data, Images, and Geospatial data (Bickler, 2021). Data scientists often find it challenging to process archaeological data, as it can be very slow to create where the timeframe can be from a few years to a few decades (Bickler, 2021). The resulting database despite the long timeframe can be classified as massive chunks of complex contextualized information (Bickler, 2021). Therefore, ML has been a transformative technique to clean, process, create models, and interpret subsequent data in the field of Archaeology. In the past, there have been different approaches for sex classification using cranial traits and the femur, for example, and examining the performance of univariate and multivariate techniques. These techniques include k – nearest neighbor, binary logistic regression (BLR), linear (LDA), and quadratic (QDA) discriminant analysis. Previously published papers have applied the same classification framework rather than applying individual classification rules which is much more comparable, and useful for researchers and scientists analyzing this community Hörr et al. (2014). As mentioned earlier, most studies on classifying sex use the human pelvis or cranium but it's proven that they lack "generalisability" and absence of solid testing framework (Del Bove and Veneziano, 2022). Human variability is only used in a limited amount during the training phase of machine learning applications. Del Bove & Veneziano (2022) draw an example by saying individual populations, or groups of populations, have been prioritized due to the difficulty of accessing worldwide data on genders. A further problem with

testing is that it is often performed on too small a sample to be statistically reliant, and the samples used during the training are often from the same set of data, which may limit its ability to be applied population- inclusively (Del Bove and Veneziano, 2022). Source and sample size are not the only limitations of previous approaches. Considering a large number of variables can also be problematic. As a result of morphological integration, cranial variables are highly correlated, and this can make the dataset redundant. When there are more variables, there is more redundancy, which increases the risk of overfitting a model (Bove, Veneziano, 2022).

At present, archaeology faces the challenge of working with unsupervised data, meaning it is unknown to what class an artefact belongs, and that classification must be determined by data mining (Horr et al., 2014). In metric sex estimation, discriminant analysis and machine learning algorithms, such as artificial neural networks (ANN) and naïve Bayes classification (NBC), are also used, where measurements are used to measure both size and shape differences. In the past, researchers have observed that both QDA and K-nearest neighbors did not fare well when it came to the sex bias criterion. This was due to the fact that the discriminant functions they produced resulted in a higher misclassification rate for one gender compared to the other. BLR seemed to be reliable to most data scientists since it depends on making fewer assumptions than LDA (Nikita, Nikitas, 2019). The comparisons between ANN and more traditional approaches such as LDA, QDA, and BLR have been made, and it was determined that ANN performs better than the other methods, but these comparisons only apply to metric and not ordinal data (Nikita, Nikitas, 2019).

Before exploring the classification methods in depth, it is crucial to calculate the Heel Ball Index (HBI) and Breadth at Ball Index (BBI). For Naïve Bayes methods with Gaussian distribution assumptions to yield correct predictions, it becomes important to comprehend the relevance of these calculated indices. By incorporating these indices, it will enhance the model's capacity to reflect variations in foot shape and proportions associated with sex. The HBI represents the ratio between the rearfoot to forefoot while BBI represents the foot width at the ball of the foot. The models can better reflect the differences in foot proportions between males and females by taking these characteristics into account, which becomes a key component of sexual dimorphism.

3. Material and Methods

3.1 Sample collection and data preparation

The study was conducted on random European and Ghanaian populations since they were the only foot measurement data that was available. The total data set consisted of 412 rows and 12 columns. All the samples were taken using quick–drying duplicating ink uniformly spread on glass. The participants put their feet on the glass plate with normal force and then placed them on the A4-sized white paper and lifted their feet without moving the paper. Our dataset has two main sources, the European data was shared by Dr. Trudi Buck from the Archaeology department at Durham University, and the Ghanian footprint measurements were acquired from "Determination of Sex from Footprint Dimensions in a Ghanian Population" (Ableduet AL, 2015). Due to its geographic proximity and interactions across the Mediterranean, the Roman Empire historically saw a large influx of Northern African slaves in its vast expanse and diversity (Harris, 1980). Considering the historical context, the foot morphologies of individuals from places such as Ghana might provide insights that can be applied to ancient European populations, even if they are not directly comparable to them. Although we aren't making direct comparisons, Ghanaian data adds a dimension of Afro-European interaction, giving our research depth.

However, there was a notable limitation in our dataset which was the lack of measurements for boys under the age of 17. It is important to stress that this omission was not due to an oversight in our analysis but a limitation arising from the scarcity of available data for this specific age bracket (10 - 16-year-old boys). This is acknowledged and we are fully aware of the potential bias it may introduce especially for shoes with insole lengths that are below 190mm. It is possible that such bias could skew our interpretations regarding the presence or absence of certain age demographics in the fort. Our research would have been better enriched if we had collected data on younger boys, but its absence does not undermine our key goals for predicting sex for adults. We rigorously examined the available data to ensure that our conclusions were as unbiased and accurate as possible, even though they were based on the available data.

Our first step to analyse foot measurements across diverse populations, our focus steered toward different parameters in the foot dataset. The subjects in this study were perfectly healthy and free from any symptomatic deformity of the foot. The following measurements were taken on both data sets -

T1 - Length measurement is taken from the Baseline to the most anterior part of toe 1 (Foot length)

Breadth at ball (BAB) – Measurement between the most lateral and the most medial projecting points of the footprint margin at the ball (which corresponds to the most prominent areas of the metatarsal-phalangeal joints).

Breadth at Heel (BAH) – Measured as the widest distance across the heel.

The indices for both BAB and BAH were calculated using the following formula –

BBI: ball breadth index = (ball breadth/length of T1)×100

HBI: heel breadth index = (heel breadth/length of T1) ×100

Figure 1 - Diagram of a footprint indicating all measurements

All the rows with missing values were checked and removed in both the European and Ghanaian datasets and a new column "Data Origin" was added for the purpose of analysing each foot individually and identifying the origins of our outliers. After cleaning up the data, we had information on 362 people, including 170 males and 192 females. Our study participants were mostly between the ages of 20 and 30 years old, followed by those between 10 and 20. The maximum age in our European data set is a Male aged 78 and in the Ghanaian dataset, it is a female aged 35. The minimum age in the European dataset is a female aged 12 and in the Ghanaian data, it is a female aged 19. The lack of older individuals in the Ghanaian dataset is considered as a limitation in this study but this

does not undermine the integrity of our research. Additionally, Ancient societies also characterized around 20-40 years old as a time of peak physical activity, mobility, and social engagement. In context, younger adults in this age bracket or even slightly above would have made up the majority of military members, merchants, and artisans. The tablets and letters found in Vindolanda provide us with a glimpse into the demographic present in and around the fort (Stanley, 2020). The letters were found in various locations, such as kitchens, workshops, and both highly ranked officers and regular soldier quarters indicating there was a lot of communication and were mostly frequented by young adults working on a variety of chores, and responsibilities (Stanley,2020). Hence, this historical context helps to understand why our training data set skews towards younger adults between the ages of 20-40 years old.

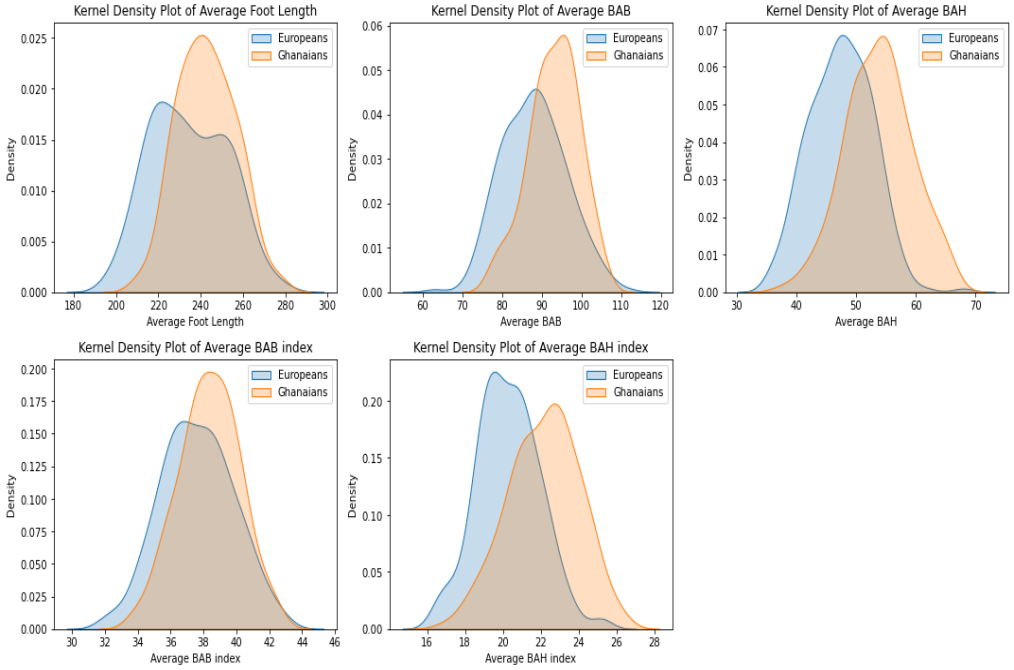
Paired t-tests were conducted for both datasets sex-wise to compare the means of corresponding left and right foot measurements across all the foot measurement variables such as Foot Length, BAB, BAH, and both of their indices. This Paired T–Test is an important statistical method for our analysis, as it compares both foot measurements, and it accounts for the dependency between the left and right foot of the same individual (Ca et al., 2020). We considered multiple factors when averaging the measurements of the left and right feet.

Our Paired T-Test for both datasets indicated that there is a significant difference statistically for European males' left foot length vs right foot length and left BAB vs Right BAB, with p-values 0.048 and 2.22e—06 (p-value <0.05) and other metrics showed no significant differences. The P-values for European females and Ghanaian males and females were above 0.05, indicating that there was no significant difference between left-foot and right-foot measurements. After observing the P-values in both sexes in both datasets, the decision to average was considered. This way it compensates for any slight variations that may exist between the left and right foot, neutralizing the effects of their possible symmetry. By choosing one foot over the other, we avoid the complexities and biases associated with choosing the wrong foot.

Consequently, the findings will be more generalizable and reliable as derived metrics reflect the underlying population characteristics. Since this study is not looking to distinguish right and left feet, but rather to understand and model sex-related differences in foot morphology. This approach aligns more closely with the research question without being distracted by specific variations that are not the study's objective. During this decision–making process, the Paired T-Tests provided a crucial guide, ensuring that the methodology was based on sound statistical reasoning and thus helping us to enhance both the validity and applicability of the study. It was necessary to average the paired measurements, as it allows condensing the data into a unified value, and we integrate the information from both feet simultaneously minimizing the complexity of the dataset. This reduction in

dimensionality does not sacrifice important information but rather simplifies our data, creating a foundation for more economical models that will be efficient and insightful.

Figure 2 - Density plots for all variables between Europeans and Ghanaians.



Average Foot Length: The average foot length reveals that they have similar distributions between them. Due to the substantial overlap between measurements between the two groups, foot length alone might not be a highly effective determinant of sex. Although it is important to consider foot length in conjunction with other variables for more accurate predictions, even though it may not be the best predictor on its own.

Average BAB and BAH: Both populations have similar distributions and suggest that the width of the foot at the ball region is ideal, with considerable overlap in measurements. Europeans tend to have a wider range of BAB values but Ghanaians have slightly higher average BAB values. Ghanaians display border dimensions compared to Europeans. They reveal steeper peaks in both categories, indicating they have stronger uniformity in this measurement.

Average BAB and BAH indices: BAB and BAH indexes tend to be marginally higher for Ghanaians, but lower for Europeans. Both indices have relatively similar distribution spreads, hinting at similar foot proportionality between the two populations based on the distribution spread of their density curves.

In contrast to Europeans, Ghanaians consistently show wider foot dimensions than Europeans. However, when we consider the proportionality indices, the distinctions become less pronounced, showcasing areas of overlap between the groups. To distinguish between male and female footprints

on Roman shoes, observable variations in foot measurements serve as an essential reference point. It increases confidence in the validity of utilizing those specific measurements for sex prediction since there is consistency in measurements between sexes across two varied groups. By leveraging both datasets, the predictive models can be trained on a wider range of foot proportions which will improve its accuracy and generalizability.

3.1.1 Difference between Europeans and Ghanaians

After cleaning the data and visualizing the distributions of European and Ghanaian foot measurements, we are going to aim to statistically validate the differences between the two groups. For this purpose, we designed a Python function, 'compare_datasets' to get the descriptive statistics of every column which are Average Foot Length, Average BAB, Average BAH, Average BAB index, and Average BAH index. For all these columns, the mean and standard deviation are calculated which provides the spread and structure of the data. T-tests were also executed to find out the statistical difference between these two datasets for each column. This is done by using the package 'stats.ttest_ind' which employs at two–sample independent T-test and compares the means from the two independent groups to observe if they are different. The results of this will demonstrate the differences in foot morphologies and the standard deviations can inform us about the consistency and range of foot dimensions with each population group. Research indicates that certain populations from different geographical locations tend to have variations in foot structure and are not just influenced by genetic predispositions but are also shaped by cultural, environmental and activity-related factors (Hoey et al., 2022). For our analysis, this becomes the primary step since the results will enable us to validate and make assumptions.

3.1.2 Cleaning our Shoe dataset

In order to conduct an analysis, it is imperative that the data be relevant and in the correct format. The shoe data had to be cleaned meticulously before being used in our study on foot dimensions. The original shoe data set consisted of a lot of Archaeological data such as sole shape, toe shape, stud pattern, thong pattern, start date, end data, and much more. While these do hold a significant amount of archaeological value it was not relevant for our analysis. Hence, these columns were removed since our primary focus was on the actual shoe measurements such as Insole length, Insole BAB, insole BAB, their indices, and Possible Owner(sex). They remain the key variables for our models. As mentioned in the background, in the past, archaeologists like Van Driel–Murray and Greene predicted sex by measuring these Roman shoes and investigating their wear patterns, material types, and style of footwear. (Greene, 2014). This approach mostly depends on qualitative analysis and cultural analogies. Conventionally this method relies on historical documentation and archaeological site

context. On the other hand, our approach to shoe data is more quantitative, based on algorithms and building predictive models. We can discern patterns hidden in vast datasets by using statistical methods and machine learning algorithms.

The initial dataset consisted of 418 rows but after removing the rows with missing measurements we had measurements of 103 shoes. This big reduction was due to the large number of missing values in the columns we need for our model. The missing measurements were also because of the fragile state the shoes were discovered and they were not in pristine conditions for getting accurate measurements. Therefore, the decision to remove these data entries was taken.

3.1.3 Calculating Stature between modern populations and Romans using footprints and shoe measurements-

Stature estimation is a key component in anthropological studies, providing critical information about ancient population's habits, health and socioeconomic conditions. Studies show a decline in the mean population stature during the Roman period both in Britain and Gaul (Redfern et al., 2015). However, this decline is not uniform because there were significant discrepancies between rural and urban populations (Redfern et al., 2015). Vindolanda Fort appears to have a distinct stature profile due to its location and purpose. Soldiers who had been trained for physical tasks were likely to be taller and the locals and slaves may have been shorter due to factors like nutrition and health (Redfern et al., 2015).

Estimating stature from osteological measures, mostly regression has been the backbone (Czibula et al, 2016). However, the development of machine learning gives a chance to improve the precision of these predictions (Czibula et al, 2016). The concept of stature has served as more than just a biological indicator throughout history. Therefore, the mean height of the population can hint at its overall well-being and nutrition over time (Czibula et al, 2016).

By analysing and comparing our footprint dataset and shoe dataset we can gain a better understating of the differences in stature between Romans and urban population. This step is crucial to observe how various factors have influenced these two demographics. We will calculate the stature and get the mean from modern footprints and the shoe data. The calculation of stature is strongly reliant on the anthropometric constants incorporated into our model which can be found in our Python code in the results. A linear regression-based method was employed for this purpose. To generate the statute, the foot length was multiplied by a predetermined anthropometric factor of 6.6 and incremented by a constant value of 60. The following formula was used to calculate the stature:

$$EstimatedStature = (Foot - metric \times factor) + constant$$

3.1.4 Dataset Alignment -

As a starting point, the footprint dataset was used to sample the males and females. As the analysis proceeded, we found significant discrepancies in the distribution of foot length based on sex. Males exhibited an average foot length of 248.96mm with values mostly around this mean and had a smaller standard deviation whereas, the females were around 225.7mm and had a slightly higher variability in foot lengths was observed. Within the focus of our study, this result suggests that the predictive models that were constructed from this dataset would develop a bias by associating longer foot lengths predominantly with males and shorter foot lengths with females. This would have a huge impact on the model's accuracy particularly when predicting the foot lengths lying close to the distribution extremes. To compare the distribution between males and females, we conducted a statistical analysis to trim our dataset, skip the class imbalance and try our best to align all foot attributes equally between males and females. However, this alignment in all the variables in our data created a new challenge While comparing the footprint dataset with our shoe data, the average foot length in the shoe dataset was greater than the shoe data set. Although, the range of the shoe dataset (130 mm-270mm) was broader than the footprint dataset (195mm -280mm). The difference in foot length between these two datasets has the potential to introduce bias in our predictions. This could affect the reliability of our models. A balance was applied to the data to minimize this inherent bias and to create a more representative sample.

Table 1. Descriptive statistics of males and females in the foot measurement data.

Features	Sex	Count	Mean	Std. Dev.	Min	25%	Median	75%	Max
Average Foot Length	Male	163	248.963	10.331	218.5	241.25	250.0	257.5	269.5
	Female	192	225.733	12.883	195.0	216.5	225.0	233.625	262.0
Average BAB	Male	163	94.822	5.631	80.0	90.5	95.0	98.0	111.5
	Female	192	85.079	6.462	63.0	80.5	85.5	89.5	101.6
Average BAH	Male	163	52.837	5.077	42.5	49.0	52.5	56.0	68.0
	Female	192	46.665	5.324	35.5	42.5	46.25	50.5	61.5
Average BAB index	Male	163	38.111	2.048	32.615	36.613	38.110	39.475	43.100
	Female	192	37.706	2.125	31.910	36.171	37.748	39.125	42.470
Average BAH index	Male	163	21.234	1.942	16.35	19.903	21.03	22.435	26.16
	Female	192	20.661	1.900	16.59	19.314	20.728	22.069	25.88

Our analysis concluded, there was a slight class imbalance between male and female samples. The male dataset consisted of 163 observations while the female dataset had 192. This is a very crucial observation since this could lead to a biased analysis. The predictive models might be inclined to predict the class with more samples due to its overrepresentation in our dataset which could lead to misleading perception of precision or recall. To address this problem, we could consider strategies like oversampling the smaller class or downsampling the larger class. The difference in mean values indicates males on average, have long feet, larger BAB and larger BAH. This is in accordance with general anthropological observations that men tend to be larger than women in height.

This raw data still contains outliers and extreme values which can affect the quality of our analysis. Therefore, Interquartile Range (IQR) is introduced to address this challenge and part of our data cleaning. An IQR represents the middle 50% of the data, the difference between the third quartile and the first quartile. There is less probability of outliers influencing this range of data than if they impacted the entire range. We're focusing on this interval to make sure that extreme values didn't distort our analysis by focusing on the bulk of the data. The mathematical formula is represented as:

$$IQR = Q3 - Q1$$

The potential outliers lower and upper bounds can be calculating using the formula:

$$Lower\ Bound = Q1 - 1.5 \times IQR$$

$$Upper\ Bound = Q3 + 1.5 \times IQR$$

This is done to Average Foot Length, Average BAB, and Average BAH. We are setting the threshold values to 218.5 for males and 262 for females to promote distinct separation between male and female distributions. This threshold will help us avoid overlapping between both sexes. Similarly, the values for Average BAB were set for males (Q1) was set to 90.5 and (Q3) was set to 98.0 and for females (Q1) was set to 80.5 and (Q3) 89.5. For Average BAH, the (Q1) was set to 49.0 and (Q3) was set to 56.0, and for females (Q1) was set to 42.5 and (Q3) was set to 50.5. The IQR is calculated using the above formulas and any data that falls outside of those bounds is considered as an outlier. The code that was designed aims to remove them and to provide a more typical and cleaner representation of the data.

After the IQR test, the statistical results still indicated some extreme values. The data still required trimming some extremes since some female foot length data was shorter than the male foot length, so we created a function ‘trim_extremes’ setting ‘n_high’ for ‘M’ and ‘n_low’ for ‘F’. As a result, we removed 5 largest male values and 5 smallest female values in the foot length data for the overall distribution. Therefore, forming our final data set.

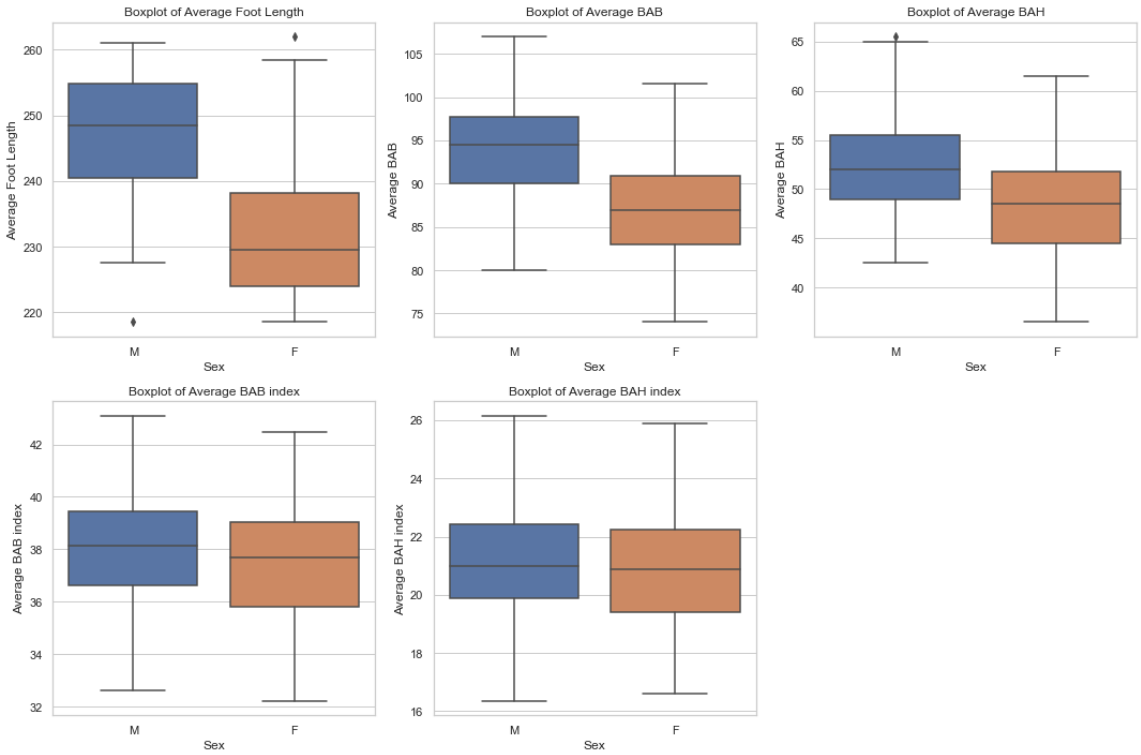
3.2 Outlier detection

After employing the IQR test to eliminate all extreme values as possible, it was important to visualize the resultant data distribution and assess the effectiveness of our outlier treatment. Thus, we used the box plot technique which will offer us an in-depth understanding of the data. All the metrics were plotted based on sex since it aligns with our research question and becomes a necessity for our analysis. Below in Figure 3 is the box plot for all the variables by sex. Upon plotting the boxplot, we could see a data point beyond the whiskers in 'Average Foot Length'. This outlier may carry relevant information about the dataset. Identifying where they originate from can help us gain a better understanding of the potential reasons behind their existence and add meaningful context to our findings.

Additional investigation revealed that the data point emerged from the Ghanaian data, a measurement of 262 mm attributed to a 29-year-old female. Since it is just one data point, the overall impact would be minimal on a data set which comprises 260 entries (after cleaning). It's not an uncommon phenomenon for a woman to have a larger foot length. Overall, the findings suggest that foot measurements can differ significantly based on the individual's sex and geographic origin. Therefore, by investigating further we can understand the impact of these outliers in our predictive models. Although it is certain that this outlier would not make an impact in our analysis, to evaluate further, we decided to build a simple linear regression model with and without the outlier and assessed its impact on predicting the 'Average BAB index' using predictors such as 'Average Foot length', 'Average BAB' and 'Average BAH'. Splitting the data into training and testing subsets, we compared the R^2 values with and without the outliers. When we looked at its impact on our simple regression model, the R^2 value was 0.9967 with the outlier and slightly lower at 0.9960 without it, indicating its minimal but existent. The decision to retain the outlier was made to include this outlier into our dataset, so we can ensure that our study is complete, and aligned with – real-world findings.

Note that, we are aware that the behaviour of data in regression models does not always precisely match data behaviour in classification models. While our simple linear regression study revealed useful insights into the outlier's effect on R^2 and this is only one dimension of our outlier evaluation. Classification tasks like our study, might react to outliers differently compared to regression. This finding is purely supplemental rather than conclusive for our analysis.

Figure 3 – Box plot of all the variables in the foot measurement data



3.3 Distribution and Normality Assessment of all the Predictors

To further understand the distributional properties of our predictor variables in our dataset, we used Quantile -Quantile (Q-Q) plots to visualize the distribution of our dataset. Many statistical tests and models rely on the assumptions of normality and is one of the crucial data pre-processing steps to confirm our final dataset meets the requirements and fits the needs of our analysis. The Q-Q plots were done for all our variables by using the 'probplot' function from the 'SciPy' library. This function compares our data quantiles to the quantiles of normal distribution providing us the opportunity to visualize the normality of each variable. The quantiles of the data were displayed on the y-axis and the quantiles of normal distribution on the x-axis. The results of these plots should have a closer alignment with the line indicating a distribution close to normal.

In our Gaussian Naïve Bayes model, it is assumed that the continuous data distribution for each class follows a normal pattern. As a result, if our continuous predictors do not follow a distribution, the underlying assumptions of our Naïve Bayes classifier may be violated. The Q-Q plots along with any noted anomalies or observations are provided in the result section of this report.

3.4 Analysis and Modelling

A structured methodology is utilized in this study for predicting the sex of Roman shoes in accordance with the conceptual framework and prior research identified in the literature review. To understand the data, we are going to be using for modelling, and identify potential challenges in the dataset, an Explanatory Data Analysis (EDA) was performed using Python before we stepped into the modelling process.

Additionally, we employed the Variance Inflation Factor (VIF) method to determine whether variables are multi-collinear. Although, we are aware that VIF is generally used in regression models, it can also be useful when it comes to Naives Bayes models and classification trees. This feature selection can be guided even in these types of classification models by understanding the relationship between different variables. This method will also allow us to know if two variables are highly correlated, they may not add much predictive power, so VIF values are useful for retaining only the most interesting variables.

The training data "balanced_foot_df" contains the footprint of the modern population and the testing data "Cleaned_shoe_data" contains Roman shoe insole measurements. The testing data's target variable is based on archaeological predictions and it is important to note that this data is raw and some shoes had some unassigned sex ('NaN' values). Initially, these 'NaN' values were removed to assess to model's accuracy in comparison to archaeologists' classifications. Later, the models were applied to the entire dataset, including the entries with unknown sex to visualize the distribution of the predicted sexes. We investigated further to observe any patterns in the predicted sexes for these unknown entries. This was done to provide a more comprehensive contextual knowledge of the data and reveal any potential biases, gaps or consistencies within our model's prediction.

The training set consisted of 314 samples of footprints and the testing set had 103 samples including 24 unknown sexes. Due to the small sample size for the test sets, a standard 80 -2- split for validation was not feasible. Instead, a 5-fold cross-validation was used to make the most of the available data. We divided the 314 samples of the training dataset into five subsets, using each subset as a validation set. The cross-validation scores were computed, and the average score served as a preliminary evaluation of the model's performance. By this approach, we were able to evaluate the generalizability of the model across different slices of data in a more robust manner. Therefore, we ensured the selected models were not overfitting to any particular subset of the training data. This was done for both models.

Since our datasets had different names, column mapping was required as a pre-processing step. We changed the names in our training data to the corresponding variables in the shoe data. "Average Foot Length" was changed to "Insole length", "Average BAH" to "insole BAH" and other variables were

done respectively. Both datasets were labelled similarly in the same way, allowing the models to recognize the corresponding variables in both datasets.

The next crucial step is using the Standardscaler which is an important scaling method that was first fitted on the training data. This step is essential for our algorithms that are sensitive to the size of input variables. Each feature in the training dataset is given a mean and standard deviation by using the StandardScaler, which normalizes the features by taking the mean away and dividing the standard deviation. The training set is standardized using these determined values. The same calculated mean and standard deviation are applied to the test set to ensure that both datasets are maintained on a consistent scale and also prevent the model from learning any information from our testing data (Roman shoes).

The two classification models that we employed for this study are as follows:

1) **Gaussian Naïve Bayes (GNB) with different thresholds** - By adjusting the classification threshold, the model's predictions are closely aligned with those made by archaeologists. Predictions are affected by threshold changes when the model assigns a class label. Usually, a 0.5 threshold is used which means the model assigns one class if the probability is greater than or equal to 0.5., and another if the probability is less than 0. This adjustment can be advantageous and disadvantageous. It can help to align the model's predictions with archaeological predictions and be tailored to fit the sensitivity or specificity. However, this has the ability to also introduce bias and may result in misleading performance metrics.

   The model uses the formula:

$$\mathrm{P}(xi \mid sex) = \frac{1}{\sqrt{2\pi\sigma^2{}_{sex,i}}} exp\left(-\frac{(xi-\mu_{sex,i})^2}{2\sigma^2{}_{sex,i}}\right)$$

   $\mathrm{P}(xi \mid sex)$: Represents the probability of feature xi within males or females.

   xi : Represents an observation for a specific feature (Insole length, BAH)

   $\mu_{sex,i}$: The mean of feature xi for sample belonging to the sex category determined from the training dataset.

   $\sigma^2{}_{sex,i}$: The variance of feature xi for samples belonging to the sex category. This serves as an indicator of how the data points for a feature diverge from the mean.

   $\sqrt{2\pi\sigma^2{}_{sex,i}}$: Normalizes the probabilistic distribution to ensure it has an area of 1 under the curve.

   $-\frac{(xi-\mu_{sex,i})^2}{2\sigma^2{}_{sex,i}}$: Determines how much the probability decreases as xi deviates from the mean.

The posterior probabilities $P(sex \mid x1, x2, \ldots\ldots, xn)$ are then computed using Bayes Theorem. The labels are usually assigned to the classes with the highest posterior probability.

$$(Sex \mid x1, x2, \ldots, xn) \propto P(Sex) \times P(x1 \mid Sex) \times P(x2 \mid Sex) \times \cdots \times P(xn \mid Sex)$$

After calculating the posterior probabilities, we adjust the decision Threshold T based on the outcomes. After calculating the probabilities of a sample that belongs to a class 'Male'(M=1) or 'Female'(F=0) based on the shoe features, we can adjust our threshold for models to make a final decision.

$P(M = 1 \mid INSOLELENGTH, INSOLEBAH, \ldots) > T$, classify as 'M', Otherwise classify as 'F'.

If we increase the T value when the false positives are high, it makes the model conservative when it assigns the 'M' label. Similarly, we can decrease the T value when false negatives are more important, which makes our model sensitive to identifying the 'M' label. We will be experimenting with different T values to see which aligns the model's accuracy with the archaeologists' predictions.

2) **Classification trees** – Different classification trees were employed to predict sex on Roman shoes and they are: 1) Pruning, 2) Random Forest, 3) XGBoost

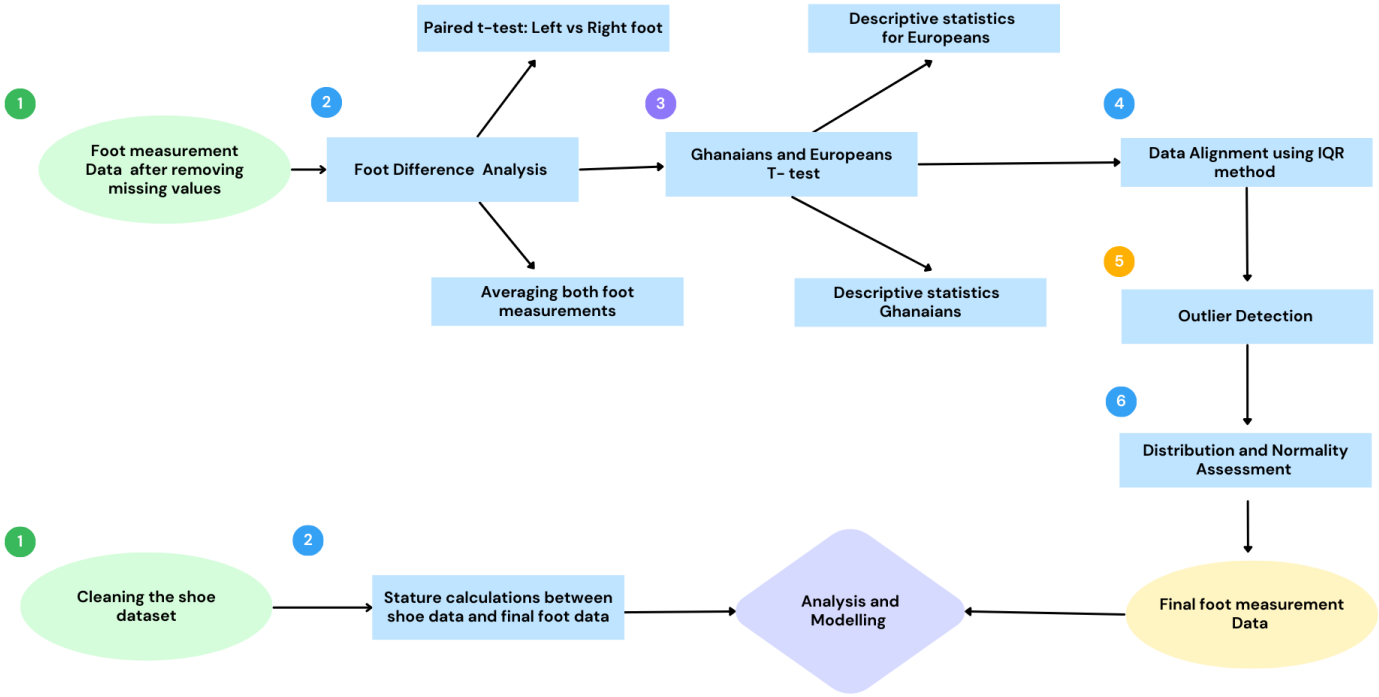
We used R studio to build the classification trees mentioned above. Similar to our GNB model we trained it on the foot measurement dataset and tested it on the shoe dataset. The metrics were evaluated using a confusion matrix and 5-fold -cross-validation for an unbiased estimation of the model's performance. The complexity parameter ('cp') was obtained and used to enhance the pruned decision tree to increase the accuracy and generalizability. The pruned tr

Our random forest model was trained with 100 trees and the training metrics were calculated. By using 5- fold- cross validation, we tuned the 'mtry' parameter to increase the optimization of the model. Later, the model's accuracy was plotted against a number of variables for splitting each tree node ('mtry'). This gave us insights into the role of adjusting the hyperparameters.

Lastly, for the XGBoost model, the training data was converted into a matrix format and the target variables were converted into binary format (F = 0, M = 1). By setting the objective to "binary:logistic" the model trained for a binary classification task using 100 boosting iterations for better enhancement. The tuning involved 5- fold- cross-validation to optimize the parameters like max depth, learning rate ('eta'), gamma, column subsampling, minimum child weight and subsample rate. We also used the early stopping feature to prevent the model from overfitting. The important variables were plotted in this model to identify which variables played an important role in classifying males and females.

Based on the evaluation metrics of all the models, the top-performing classification tree was selected and applied to the raw Roman shoe dataset to make predictions. The results of the distribution in the dataset are visualized to get an understanding of the final outcome. We conclude the section with a comparison of the model's performance highlighting all the strengths and weaknesses in context to our research goals that we aimed to achieve. The methodological flowchart can be found in figure 4.

Figure 4 - Methodological Flowchart for Analysis and Modeling



4. Analysis, Modeling, and Results

4.1 Statistical analysis between Europeans and Ghanaians

Table 2 - Results of statistical analysis between Europeans and Ghanaians

		Mean	Standard Deviation	T-test static	P -value
Europeans	Average Foot Length	234.19	18.17	-4.59	5.94e-06
Ghanaians		242.69	13.70		
Europeans	BAB	88.0	8.06	-6.02	4.27e-09
Ghanaians		92.98	6.32		
Europeans	BAH	47.46	5.25	-10.65	3.18e-23.
Ghanaians		53.76	5.55		
Europeans	BAB INDEX	37.60	2.21	-3.19	0.00156
Ghanaians		38.33	1.83		
Europeans	BAH INDEX	20.27	1.65	-9.87	1.67e-20
Ghanaians		22.15	1.85		

- ***Average foot length:*** Our mean value indicates that Ghanaian's feet are a little longer than Europeans. Although the Standard Deviation (SD) implies that there is less variation in foot length among Europeans than Ghanaians. We got a low p-value which suggests that we have strong evidence to reject the null hypothesis and there is a difference between these two populations,
- ***Average BAB:*** The mean values indicate the Ghanaians have a bigger breadth at the ball and the SD suggests that there is slightly more variability in the European data set. The t -t-static and P -values indicate that there is a difference statistically, which has evidence to reject the null hypothesis.
- ***Average BAH:*** The mean values suggest that the Ghanaians have a bigger Breadth at Heel, and the SD suggests that the variability is similar in this case. The null hypothesis is rejected for BAH as well Average BAB Index.
- ***Average BAB index and BAH indices:*** The mean values for both indicate that Ghanaians have slightly bigger BAB and BAH than Europeans. The SD suggests that there is more variability in BAB index in Europeans and Ghanaians have more variability in BAH index.

The null hypothesis is rejected between these two groups. This was a logical expectation since the indices are derived from their foundational measures.

This statistical analysis offered a quantitative lens for comprehending foot dimensions painting a clear picture of size and variety within both datasets. By calculating the mean differences and determining their statistical significance, the T-Test strengthens our analysis even further. The SD in the Ghanaian dataset is low in foot length and BAB which implies there is a higher degree of uniformity in foot dimensions compared to Europeans. The final t-tests conclude that the null hypothesis is rejected in all metrics which confirms that these differences are not random and they are statistically significant. Overall, this analysis has shed light on both populations by explaining the difference between the two groups.

4.2 Statue calculations between footprint and shoes

Table 3 - Mean of Stature calculations between Modern population and Romans

Data	Mean Estimated stature	Standard deviation
Footprint (modern population)	1.629 meters (5.3ft)	100.8
Roman Shoes	1.479 meters (4.8ft)	199

While comparing the footprint and the shoe dataset, there is a noticeable difference in both the mean estimated statures and standard deviations. The disparity provides significant insights into the characteristics of the modern people and Romans. The footprint data indicate a slightly taller and more homogenous group, presumably indicative of a population with superior dietary and health conditions, with a mean stature of 1.629 meters and a standard deviation of 100.8mm. The shoe data, on the other hand, indicates a shorter and more diverse population because of its mean stature and deviation. This may indicate a wider variety of socioeconomic, dietary or environmental factors impacting stature.

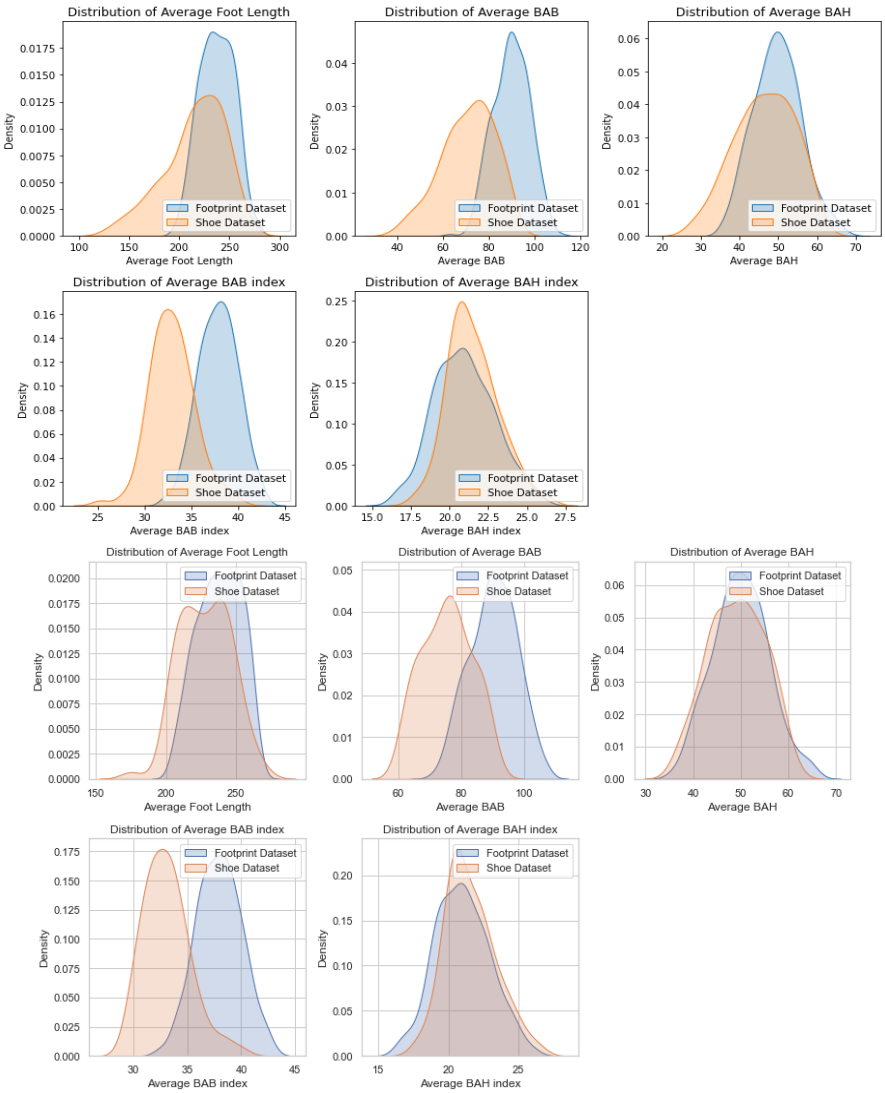
4.3 Post-Dataset Alignment

Before we cleaned and aligned the dataset, it exhibited noticeable disparities between males and females, and our primary goal was to eliminate potential sources of bias and pave the way for model training by maintaining integrity with our data-driven decision-making. The dataset consisted of 163 male samples and 192 female samples. Furthermore, the alignment of the footprint with the Roman shoe dataset (‘cleaned_shoe_data’) was required to ensure they were equitable for our model training.

In order to ensure fairness and avoid unintended bias, the datasets had to be aligned based on foot length measurements and insole length measurements, respectively. Scaling, shifting, and normalizing techniques were used in our alignment strategy. By making these adjustments, the "Shoe Dataset" was transformed statistically to correspond to our footprint dataset. Initial results showed that the dataset had a mean foot length of approximately 237.74 and a standard deviation of approximately 15.28. Meanwhile, the "Shoe Dataset" displayed a mean insole length of approximately 215.01 with a standard deviation of approximately 30.15. We bridged this gap by precisely aligning the two datasets, ensuring that foot length and insole length mean and standard deviation values were much closer, eliminating potential bias caused by original discrepancies. As mentioned in my methodology we tackled outliers using the IQR method, and a simple method of random sampling and using the 'len()' and 'min()' functions. Post these operations, both males and females were balanced and out and the final data set consisted of 157 entries for each sex. Males had a mean foot length of 248.23 and females had 227.25 after omitting extreme values and outliers. It was noteworthy to observe the distinctive distribution of the 'Average BAB' and its index. Distribution charts showed that the 'Average BAB' in the foot measurement dataset was significantly greater than that in the shoe dataset. The difference isn't due to random variation in the data. It provides an intriguing insight into how human anatomy evolved. This indicates a subtle but meaningful change in human morphology since the Roman period, as the Ball at Breadth (BAB) of feet may have increased. In addition, addressing intergroup differences, it was noted that Ghanaians had a slightly larger BAB compared with Europeans. As a result of this observation, it is crucial to consider since it will help us identify the important predictors.

To enhance clarity and transparency, we present the updated metrics for the footprint dataset post-alignment (The metrics of the updated footprint dataset can be found in Appendix 7.3). Though there was more room for manipulating the distribution of foot data to mirror the shoe data more closely, altering it further led to the loss of critical samples and the strategy did not yield better results, which led us to reconsider it. To provide a clear representation and demonstrate our progress in data alignment, we have prepared comparative plots illustrating the distribution between footprint data and shoe data. Our alignment methods are illustrated below in Figure 5 providing an intuitive glimpse into the improvements we were able to achieve.

Figure 5 - Density plot of the footprint data and shoe data after aligning both



4.4 Normality Distribution

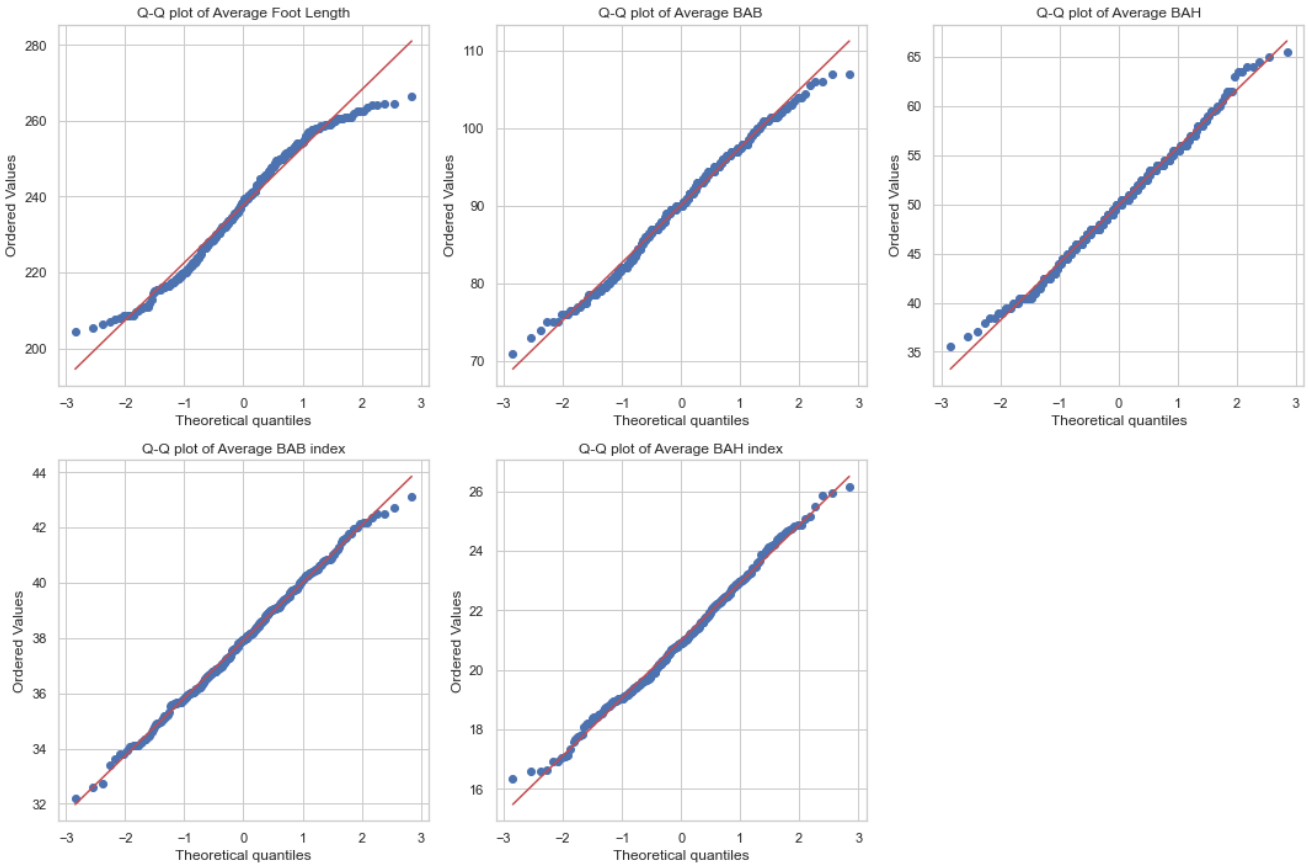
Q-Q plots -

Now that our dataset is balanced and aligned with the shoe dataset, we can venture into checking the normality using Q-Q plots and explore all our variables with Exploratory Data analysis (EDA). By this systematic approach, we can ensure that our modelling efforts are informed by an in-depth and subtle comprehension of the data environment. In Figure 6, we present the Q-Q plots of all the predictor variables serving as the foundation step on our data assessment journey.

Upon analyzing the Q-Q plots, there is a slight deviation at both tails for 'Average Foot length' in an idea plot, the data points are aligned perfectly with the reference line, implying normal distribution. However, in our scenario, we noticed a few deviations in the 'Average Foot Length'. There are

noticeable deviations in the upper and lower tails particularly around the -2 mark on the lower side and somewhere between 1 and 2 on the upper side. The data points in the middle are close to the reference line and suggest peak distribution in the centre. However, the slight rise of data points above the line at 0 reveals a minor positive skew, indicating that the upper half is slightly more distributed than its standard normal counterpart. These minor inconsistencies in 'Average foot Length, are noted and are acceptable. These slight variances were expected given the nature of our dataset. These little deviations from perfect normality are often reasonable for modeling. For 'Average BAB', the data follows closely to the reference line indicating the distribution is close to normal with a couple of data points outside the line. Similarly, 'Average BAH', 'Average BAB index', and 'Average BAH index' is close to the line and show a similar trend with deviations. All these variables except 'Average Foot Length' portray normal distributions and the slight deviations, and data points outside the line are not of such scale that would raise any concerns.

Figure 6 - Normality check using Q-Q plots for all the variables

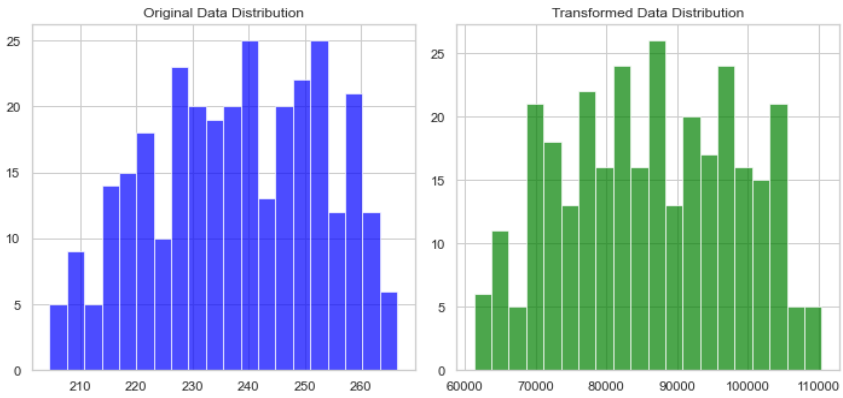


Although we did attempt to check if 'Average Foot Length' can achieve perfect normality, we utilized the Box-Cox transformation method. The fundamental idea behind this data transformation is to check an appropriate (λ) that can be applied to all data points, leading to an enhancement of the distribution's normalcy. We plotted the distributions and took a Shapiro-Wilk test to check for

statistical validation. First, Our Shapiro-Wilk test revealed our 'Average Foot length' was not perfectly distributed, the static was recorded at 0.972 and the p values were 0. After applying the Box-Cox transformation we plotted histograms (Figure 7) to see the difference between the original data and the transformed data. Another Shapiro-Wilk test on the transformed data showed that the Box-Cox transformation had limited effects. The Static was recorded at 0.973 and the p-value of 0. However, this transformation failed to produce notable improvements, and it becomes prudent to keep the data in its original form. Thus, with these considerations, we have elected not to alter the data and keep it in its original form.

This was expected given that we know the nature of our footprint data and the visually almost similar histograms further emphasized this point. Our footprint data was restricted, which presented an underlying challenge. This dataset was concise and could not be expanded further. We shifted our strategy due to our dataset's inherent limitations and negligible transformations.

Figure 7 - Distribution of foot length before and after transforming the data with Box-Cox method



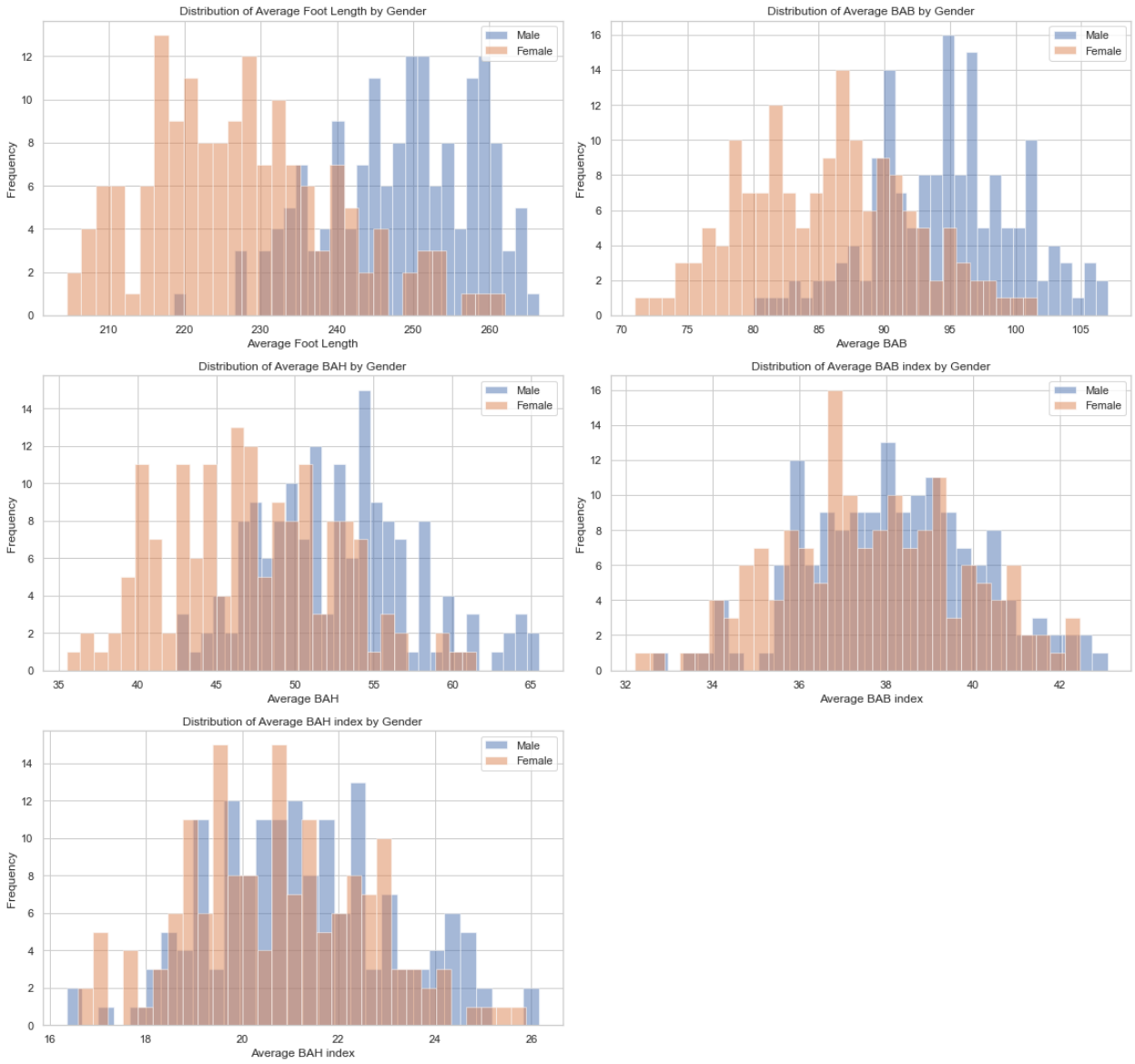
4.5 Explanatory Data Analysis

Unraveling the complexities of our dataset requires a comprehensive exploration and this is precisely where EDA becomes invaluable. We will visualize the data's underlying patterns, distributions, and relationships across all the variables. Note that, we have previously employed boxplots in our methodology to identify and visualize the outliers in the dataset. Furthermore, we will use histograms distinguishing data points by sex. This sex-based segmentation will provide us with a detailed view, and shed light on the differences in the distribution of all variables. Additionally, pair plots were utilized to highlight the relationships across variables, along with insights gained from the correlation matrix.

In Figure 9, the histograms illustrate the Anatomical variances in foot measurements between males and females. In 'Average Foot length' the males exhibit a peak in a noticeable rightward shift, demonstrating greater foot lengths compared to females. The average is seen in males which is around

250 and for females it is around 230. This observation is consistent with some anthropological investigations which typically indicate that men have larger foot sizes than women (Luo et al., 2009). Similarly, 'Average BAB' exhibits a bell shape for both sexes but they are noticeably displayed from one another. The male distribution peaks at about 94 and the female center more closely around 87. While they share some values and overlap, there is still a noticeable difference between them. The 'Average BAH' are also bell-shaped and the average male tends to have a bigger BAH than an average female. For their indices, it portrayed a more unified pattern, and the distributions closely lined up for males and females. The separation is very mildly noticed. Finally, the metrics of foot length, BAB, and BAH establish distinct boundaries influenced by sex. The indices appear more neutral suggesting less variance.

Figure 8 – Distribution of male and female in all variables

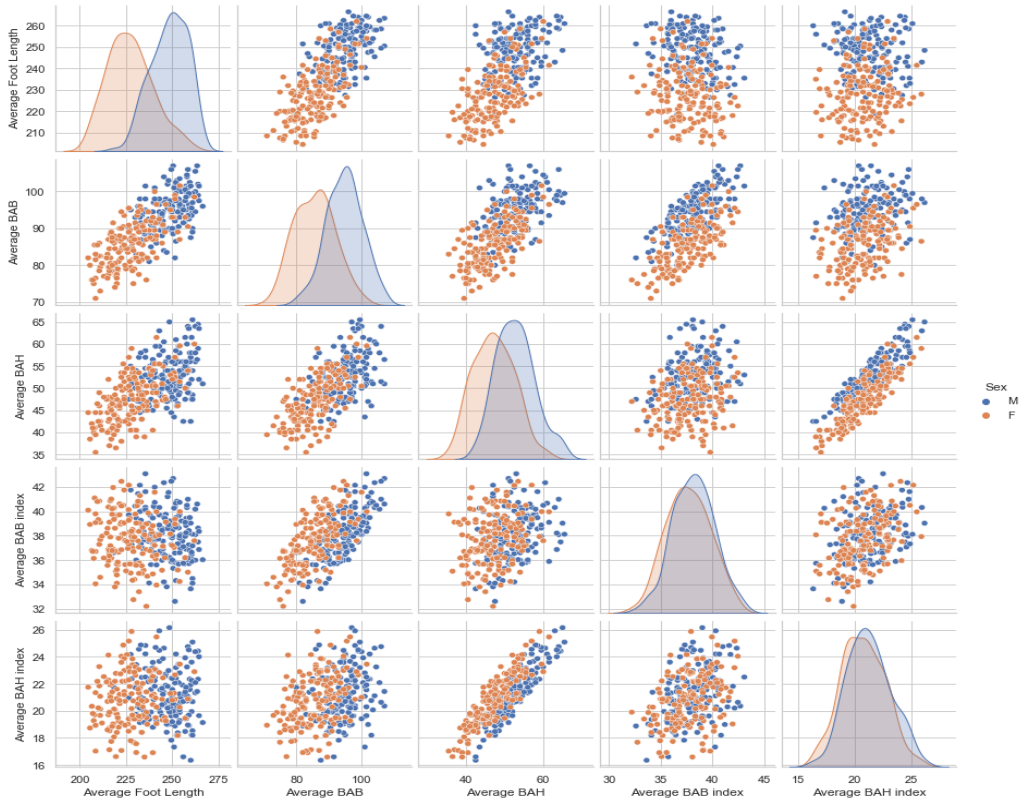


Therefore, when compared to female distributions, males generally shift to the right indicating larger measurements. It is important to note that there are still overlaps. While these histograms offer us a broad understanding, they are not our conclusive indicators. In some cases, females exhibit the same foot lengths, BAB and BAH found in males, and vice versa.

4.5.1 Correlation analysis

This plot represents the relationship between every pair of variables in our footprint dataset. Using this grid, we can determine whether two variables are linear or nonlinear or do not share a relationship. Besides providing us with a representation of bivariate relationships, they also illustrate the distribution of individual variables, providing us with an overview of both single-variable and multivariate patterns.

Figure 9 - Scatterplot matrix of footprint data

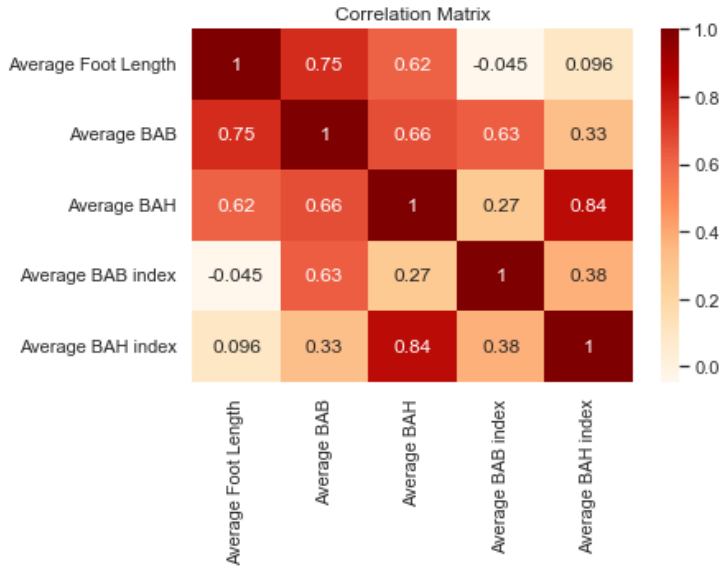


In Figure 9, the blue data points are male, and the orange data points are female, and we observed that the female data points cluster more densely compared to males. Firstly, we noticed a strong upward trajectory between ‘Average Foot Length’ and ‘Average BAB’ suggesting that as the foot length increases, BAB also tends to expand. However, females have a better correlation than males in this grid. Similarly, ‘Average Foot length’ and ‘Average BAH’ show a strong positive correlation suggesting bigger BAH with increased foot length but it is slightly weaker compared to BAB. Both

BAB and BAH also seem to exhibit a strong relationship implying that individuals with broader BAB would also have bigger BAH. Finally, observing the relationship between BAB, BAH and their indices showcased an upward trend. This is expected since the index represents the raw measurement as a percentage of the foot's overall dimensions, and it makes sense that the raw measurement increases corresponding to its index. To understand the dynamics between different parameters, particularly their strength, and interactions, we will focus on the correlation matrix.

In Figure 10, to quantify further we plotted a confusion matrix. A strong correlation of 0.75 is noted between foot length and BAB, as we noted in our scatterplot matrix. In contrast to BAB, foot length, and BAH also show a moderately positive correlation value of 0.62. The correlation value between BAB and Bah is 0.66 highlighting a moderately positive correlation value. However, this matrix adds more clarity to the indices of BAB and BAH. Average BAH is strongly associated with its index at 0.62, whereas BAH is more closely associated with its index with a value of 0.83. Both the scatterplot and correlation matrix concluded that foot length is weakly correlated with BAB and Bah indices. As a result, the use of both matrices helped us understand the foot anatomy of our dataset. By deeper exploration for sex-specific patterns, we can conduct a short correlation analysis by sex to understand the foot anatomy for males and females which will also lay us a solid foundation for our subset analysis.

Figure 10 - Correlation matrix of footprint measurements

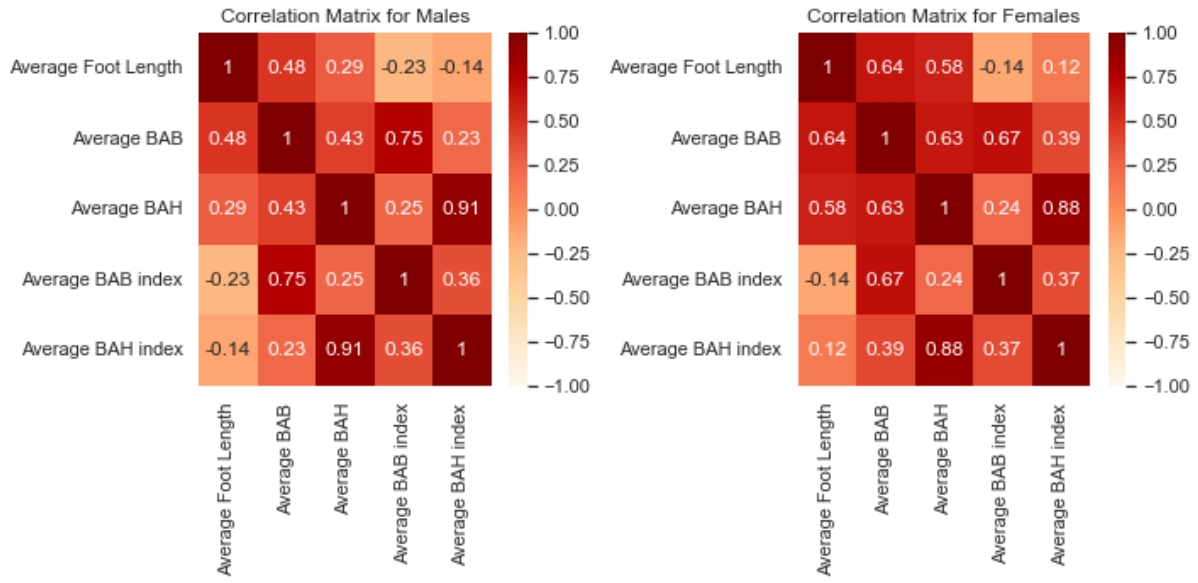


4.5.2 Correlation Matrix by Sex –

In exploring the intricate details of foot anatomy across sex, correlation matrices (Figure 11) reveal compelling distinctions. Firstly, the correlation between foot length and BAB is stronger among

females (0.636) than among males (0.477). In comparison to males, BAB increased more pronouncedly as the female foot length increases. There is also a striking correlation between BAB and BAH for females more than males, suggesting that female foot dimensions follow a consistenr proportional relationship. The correlation between Average BAH and its index is 0.88, highlighting the proportionality of heel breadth in females. Similarly, the males have a correlation value of 0.91. It is evident that specific foot measurements are positively correlated for males and females, but their levels are subtle yet profoundly different. There is a stronger association between 'Average BAH' and 'Average BAH index' in both sexes compared to 'Average BAB' and 'Average BAB index'. Based on this finding, the heel's breadth might be more important in determining foot structure, possibly because the heel is a primary point of weight bearing. Having established this foundation, we can now explore subset analysis in more depth and understand the implications of these vital parameters.

Figure 11 - Correlation matrix by sex



4.6 Finding the best Predictors for our models

Variance Inflation Factor (VIF) -

Taking a close look at the predictors had a critical influence on our ability to uncover patterns, correlations, and insights that were previously unexplored in this relationship. Before modeling, this step was essential to address multicollinearity within our predictors. Multi-collinearity could skew results and obscure genuine correlations, compromising its accuracy and credibility. There were compelling associations revealed at an earlier stage of the study. EDA showed that 'Average BAH'

was more important than 'Average BAB'. We observed different correlations across genders, suggesting that our predictors interacted differently. It was evident from these initial insights that it was imperative to select predictors carefully, making sure they were not only relevant but also independent.

The VIF method was chosen to find the multicollinearity in our training dataset. We conducted this test using all our predictors. Below are the variables with VIF values:

- Average Foot Length: 389.068763
- Average BAB: 690.850642
- Average BAH: 456.248227
- Average BAB index: 302.670398
- Average BAH index: 282.741924

Generally, any VIF value above 10 indicates high multicollinearity. In addition to exceeding this threshold by a large, margin the initial values were remarkably high. Due to these concerns, we developed a method for simplifying the predictors, focusing on preserving relevant variables and minimizing collinearity. After trying out different combinations of predictors, the VIF values gradually declined and are in acceptable form.

- Average Foot Length: 1.623664
- Average BAH: 1.623664

These values are < 2 which is not a significant concern and interestingly our VIF results mirrored our findings in our EDA. A major difference between BAH and BAB was already noted during our EDA and VIF analysis has confirmed that BAH remains a critical variable in the refined model. By aligning VIF and Eda, we demonstrate our analytical approach's robustness and reinforce our predictive model's solid foundation.

4.7 Model Results

4.7.1 Gaussian Naïve Bayes Results

Based on the outlined methodology, we trained our Gaussian Naïve Bayes (GNB) models using foot measurements and tested them using raw data and unseen shoe measurements. The combination of 'Average Foot Length', 'BAH', and 'BAH index' showed us promising performance, leading our model with better accuracy rates. Notably, we observed diminished accuracies when all the predictors were used to train and test the model. This highlighted that predictor selection is crucial for model performance. The following thresholds were used for models: 1) 0.1, 2) 0.4, 3) 0.5, 4) 0.6

In this section, we will discuss the results of our training set and demonstrate the model predictions alongside distributions from the test set without 'Nan' values in the target variable 'Possible Owner (Sex)'. Following that, we will show predictions based on applying our model to the full dataset, including 'NaN' values. The training accuracies of all the models are in the model comparison section at the end.

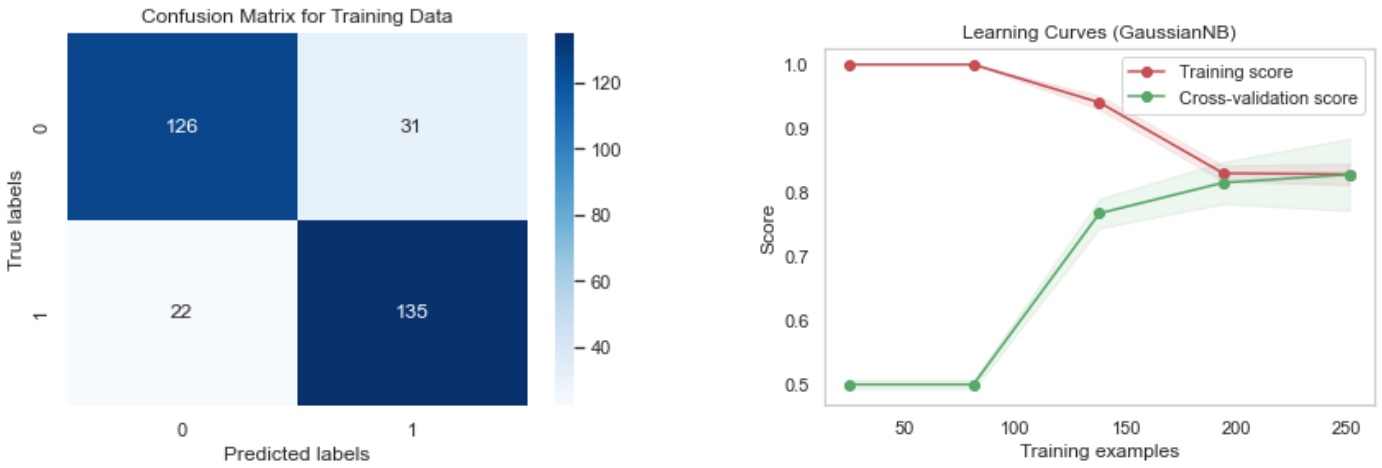
Training metrics and K fold cross-validation –

The model achieved an accuracy of 83% with precision and recall at 81% and 86%. The recall was 86% indicating the model's ability to identify the actual positive cases across all the training samples. Finally, the F1-score registers as evidence for our model's consistency during the training phase.

According to the confusion matrix (Figure 12), 135 males were correctly classified as males, while 126 females were correctly classified as females. As, a result 32 females were mistakenly classified as males (false positives), and 22 males were wrongly predicted as females (false negatives).

Based on the cross-validation score, it maintains a steady average of approximately 0.82, providing a reliable indication of its effectiveness on unseen data. We can conclude that our model does not underfit or overfit on unseen data but falls somewhere in the middle. Furthermore, we got some insights after plotting a learning curve which was obtained from using the function 'plot_learning_curve'. When the model is trained only with a few training samples, for example maybe around 20 -65, it demonstrates near-perfect performance. As a result, the model might be memorizing the training data instead of generalizing it. The low cross-validation score implies its ability to generalize well to unknown data. Gradually, however, the model becomes more general with more training data. When the model reached and trained with 201 samples, it settled at approximately 0.85 indicating a balanced performance and suggesting that inclusion of more samples decreases performance. Overall, the GNB model has proven to be capable of learning from available training data, transforming from overfitting to more reliable performance.

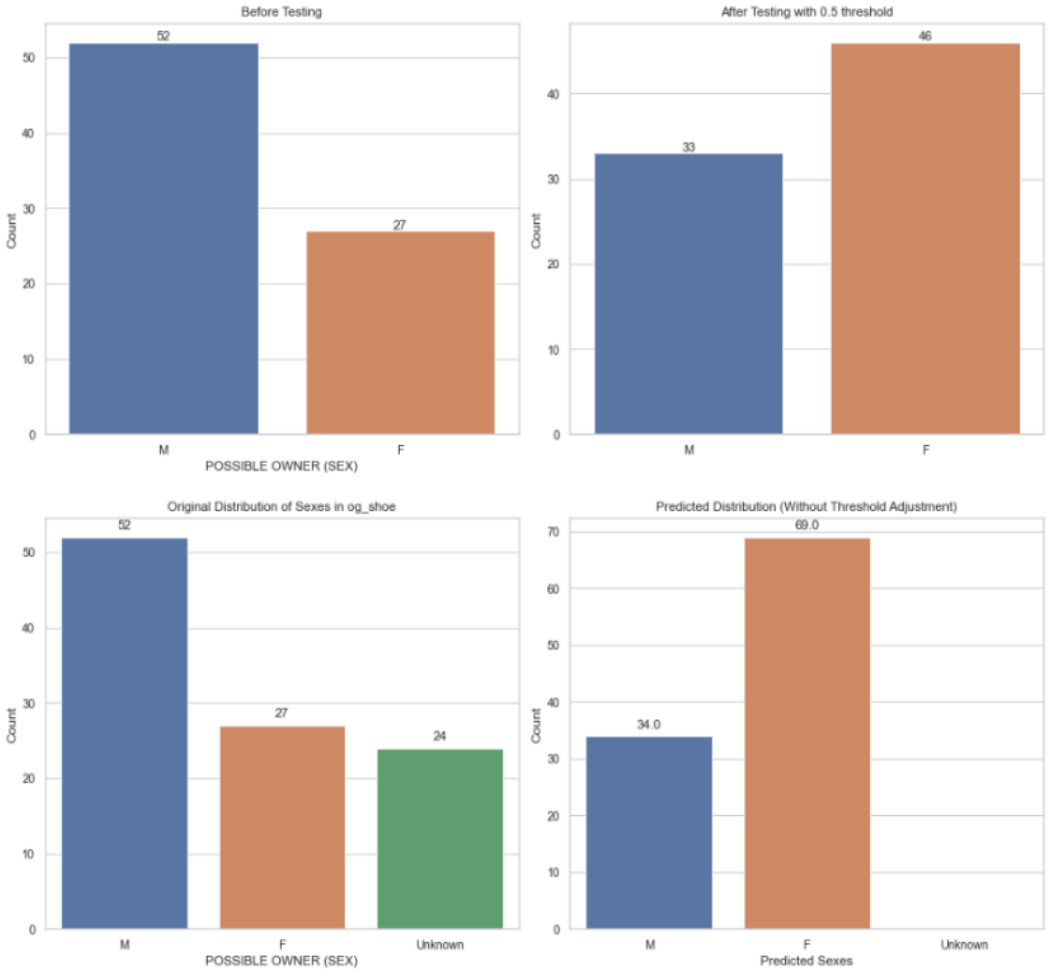
Figure 12 - Confusion matrix and learning curve of the GNB model



Below is a detailed comparison of the results of our trained model and the distribution from our Test data:

1. *Predictions with a Threshold of 0.5 (Normal)*

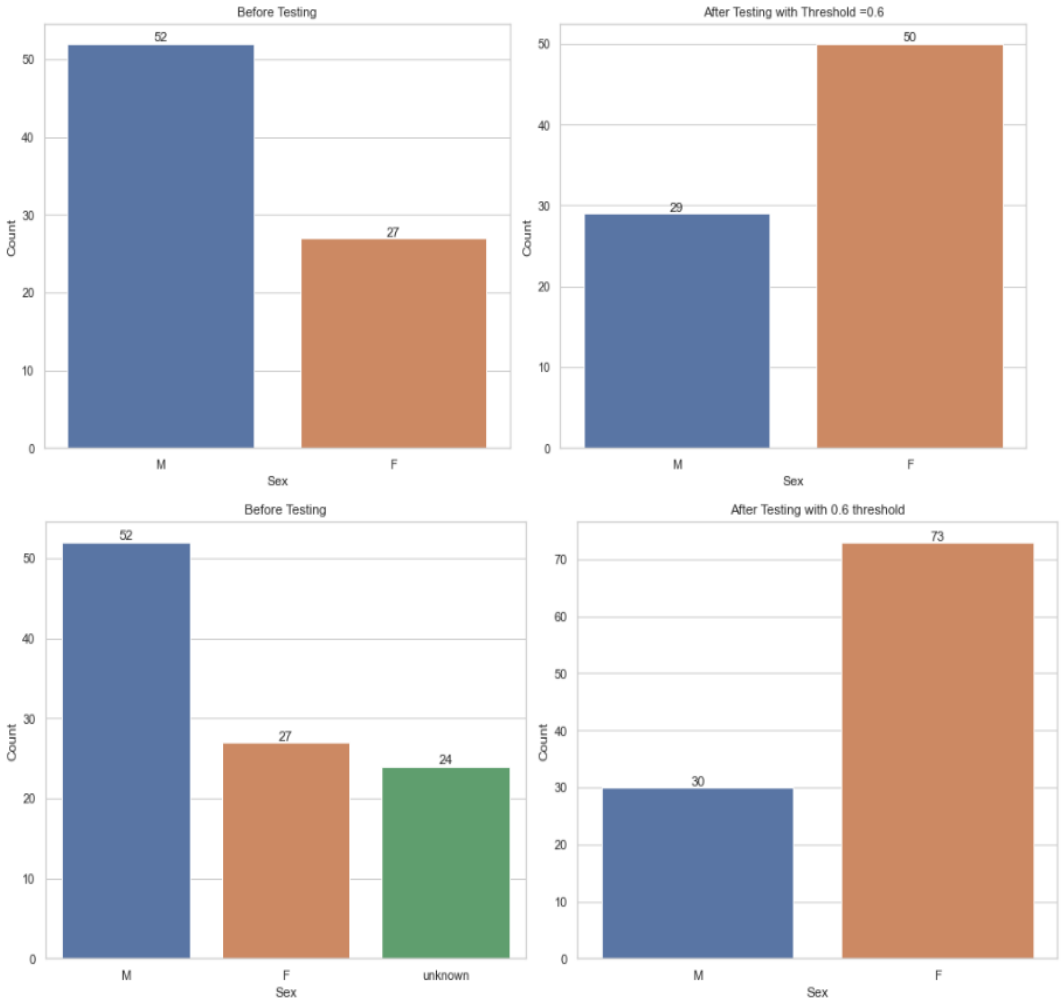
Figure 13 – Predictions with 0.5 threshold. Top – before and after without NaN values. Bottom – When trained models are applied to the Whole dataset with NaN values.



The trained model produced distinct observations when applied to the shoe dataset with 0.5 as the standard decision threshold. In the dataset without NaN value, the model predicted 33 males and 46 females. Compared to the dataset with NaN values, the model predicted 34 males and 69 females. Clearly, the difference in these distributions suggests that NaN values play a role in the model's predictions, emphasizing the importance of quality and completeness of data. It appears that the model at 0.5 threshold tends to lean slightly towards classifying samples as females, especially for uncertain or borderline samples.

2. *Predictions with a Threshold of 0.6*

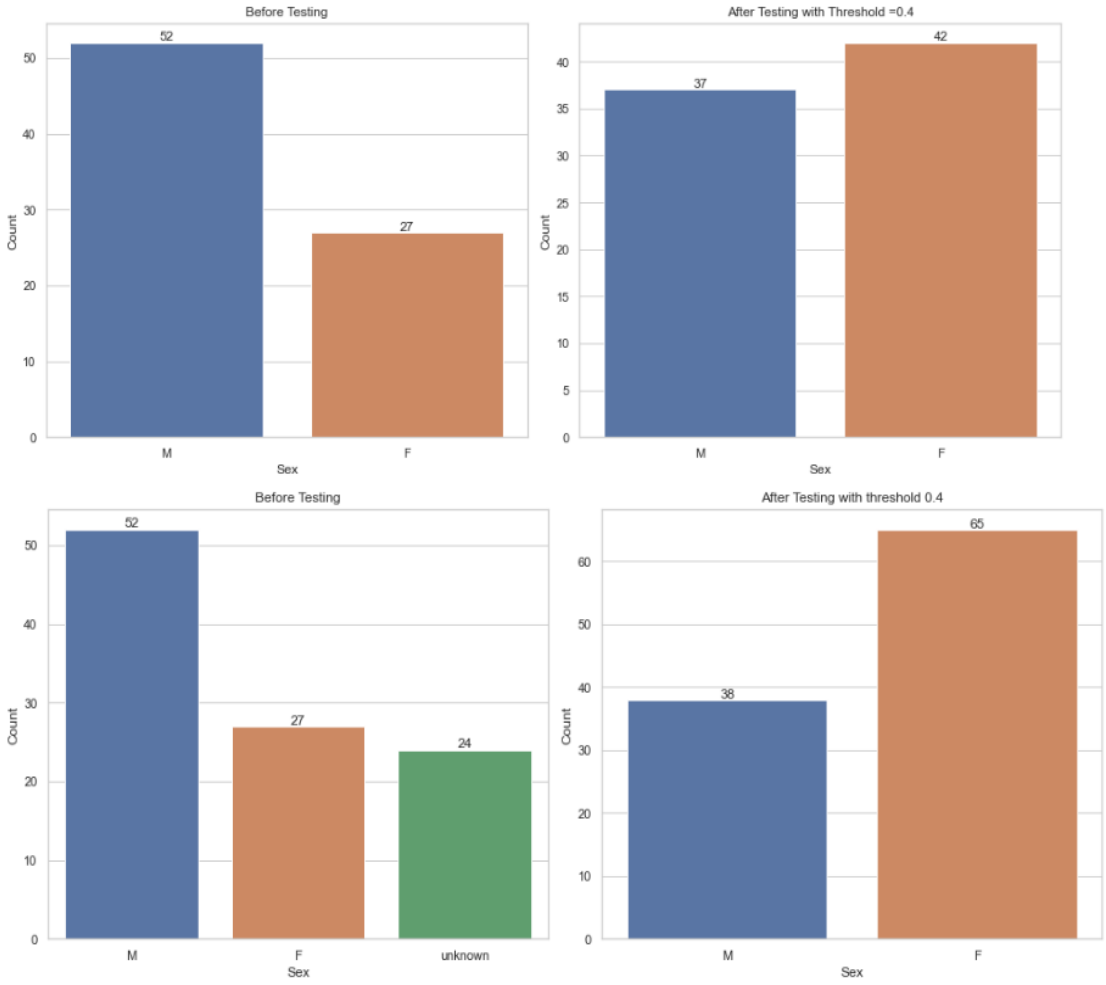
Figure 14 - Predictions with 0.6 threshold. Top – before and after without NaN values. Bottom – When trained models are applied to the Whole dataset with NaN values.



In both datasets, female predictions have increased and male predictions have declined. The ratio of male predictions to female predictions increased from 46 to 50 in the dataset with NaN values. Female predictions jumped from 69 to 73 in the dataset with NaN values, whereas male predictions dropped from 34 to 30. It highlights the model's enhanced sensitivity to a threshold, leaning towards predicting more samples as females under a higher threshold.

3. *Predictions with a Threshold of 0.4*

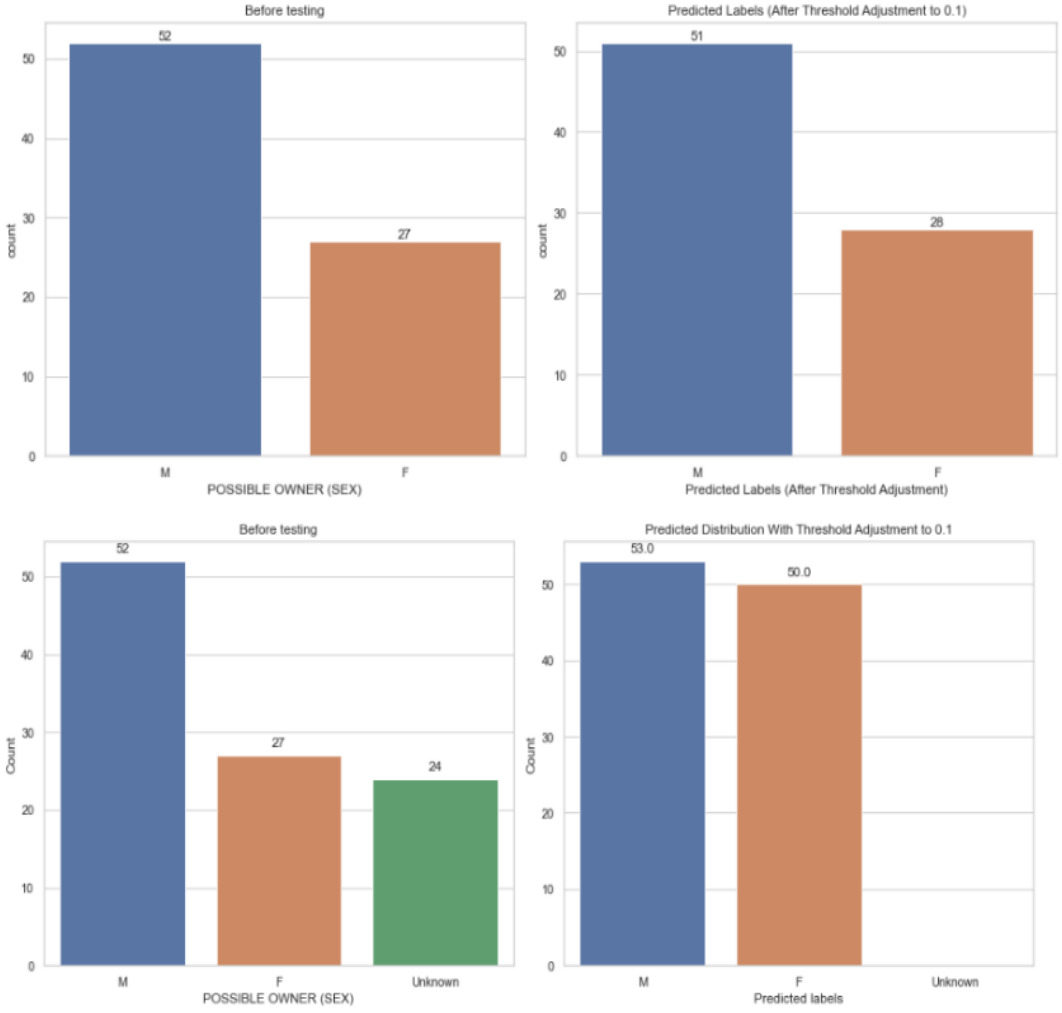
Figure 15 - Predictions with 0.4 threshold. Top – before and after without NaN values. Bottom – When trained models are applied to the Whole dataset with NaN values



A smaller threshold of 0.4 resulted in a shift in predictions. The model predicted 37 males and 43 females in the test set without NaN values. In comparison to the initial threshold of 0.5, male predictions increased from 33 to 37 and female predictions decreased from 36 to 42. Female predictions dropped slightly from 69 to 65 for the test set with NaN values. An adjustment below the threshold of 0.5 leans the model towards more males. Changing the threshold directly impacts how the model differentiates between male and female predictions when it comes to sensitivity and specificity.

4. *Predictions with a Threshold of 0.1*

Figure 16 - Predictions with 0.1 threshold. Top – before and after without NaN values. Bottom – When trained models are applied to the Whole dataset with NaN values



After adjusting the threshold to 0.1, the model's predictions have come very close to the initial distribution. Based on the distribution of our original test set which consisted of 52 males and 27 females, the model has predicted 51 males and 28 females. Hence, at this threshold, the model's behaviour aligns with the data's inherent distribution, indicating a good fit. In the test set with NaN values, 53 males and 50 females were predicted by the model. We were able to match the archaeologist's predictions by adjusting to this particular threshold.

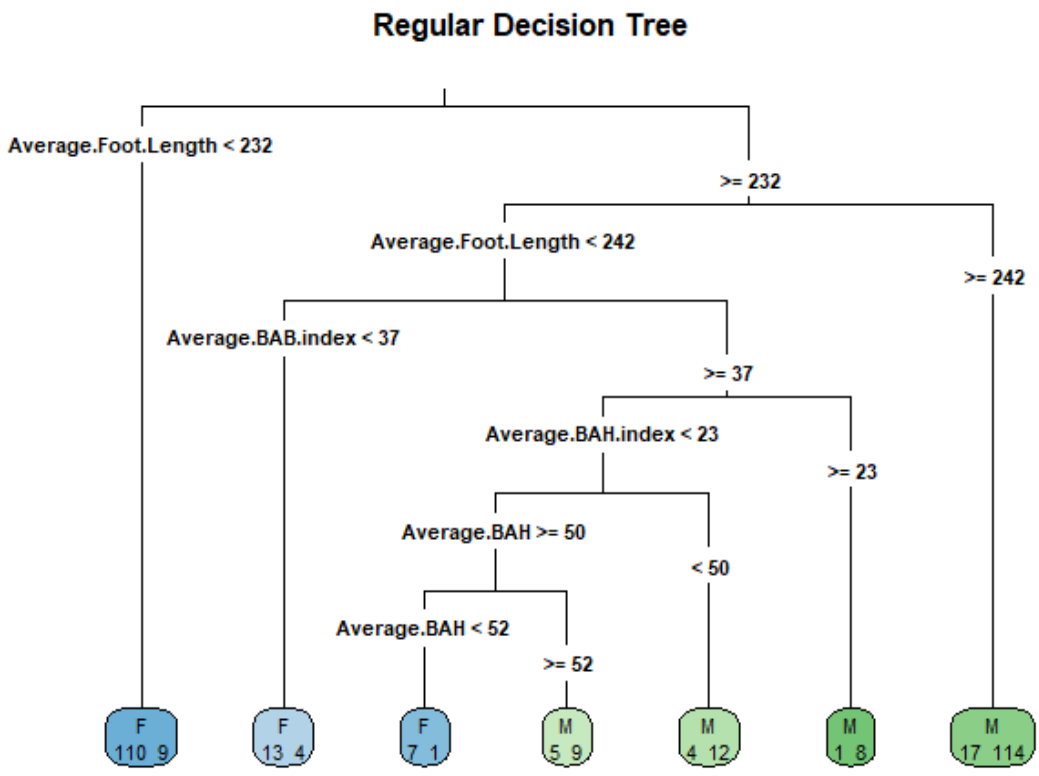
4.7.2 Classification trees –

In addition to our exploration of GB models with different thresholds, we examined classification trees to refine our prediction further. Every trained model was trained on the foot measurement dataset and then evaluated on unseen shoe data after removing the rows with Nan values in the 'possible owner (sex)' column. Following these first tests and evaluations, the model with the greatest accuracy was chosen and applied to the shoe data with Nan values to predict the sex. The resulting predictions were visually distributed providing us with insights into the possible sex distribution in the unlabeled data.

Compared with GNB's probabilistic approach, classification trees offer us hierarchical decisions and are presented graphically. This provided a distinct alternative. The following tree-based models were implemented: 1. Pruning, 2. Random Forest, 3. XGBoost

The training dataset was fitted using a standard classification tree. The tree had seven terminal nodes and was built using all the variables. The unpruned classification tree in Figure 17 outlines its structures and its decisions.

Figure 17 - Visualisation of Unpruned Classification Tree Model



Most of the data is primarily classified based on one variable 'Average foot Length'. These initial differences are heavily influenced by this variable. The majority of people with foot lengths less than 232mm tend to be female. As foot length increases beyond 242mm, the chances of an individual being male also increase. However, between the range 232 and 242, there are other variables such as 'Average BAB' and 'Average BAB index' that play a role in determining the sex. Both of these measurements together with foot length, capture numerous dimensions and proportions of the foot, bolstering the idea that foot morphology can be a powerful indication of sex.

Before pruning the tree, we had to analyze the effectiveness of our decision tree model, we used 5-fold cross-validation on the training dataset which consisted of 314 samples. This was used to get an unbiased estimate of the model's assessment metrics and to reduce our chances of overfitting, ensuring that our model generalizes when it comes to new or unknown data.

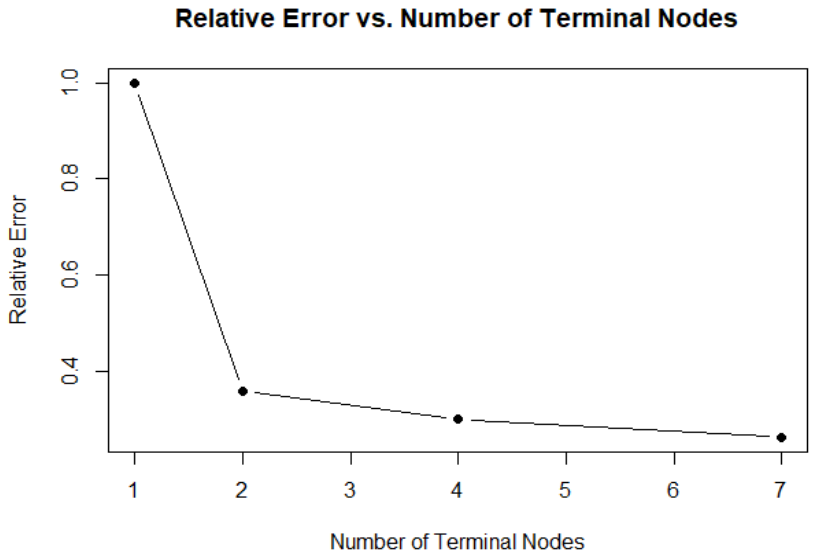
To determine the size and complexity of the final tree, the model was evaluated across three different complexity parameters (cp). The results are as follows:

Table 4 – Cross-validation results for different Complexity parameters

Cp value = 0.01273885	
Accuracy	80.31%
Kappa	61.82%
Cp value = 0.02866242	
Accuracy	80.28%
Kappa	60.49%
Cp value = 0.64331210	
Accuracy	55%
Kappa	10.63%

We constructed a plot (Figure 18) that displays the relative error versus the number of nodes. With each subsequent split, the relative error decreases, demonstrating the effectiveness of the decision tree in partitioning the data. First, there is no division at the initial node and provides a broad prediction for all data points since it is the only node that is present. This gives us a 100% error rate demonstrating there is no class distinction. The first split notices a significant drop in relative error around 35%. This emphasizes the importance of the tree's initial choice capturing prominent patterns in our data. Finally, the error rate dropped around 29% and 26% after the tree branches to four terminal nodes. It appears that every addition split refines the predictions a little less than the previous split, due to the diminishing reduction in error.

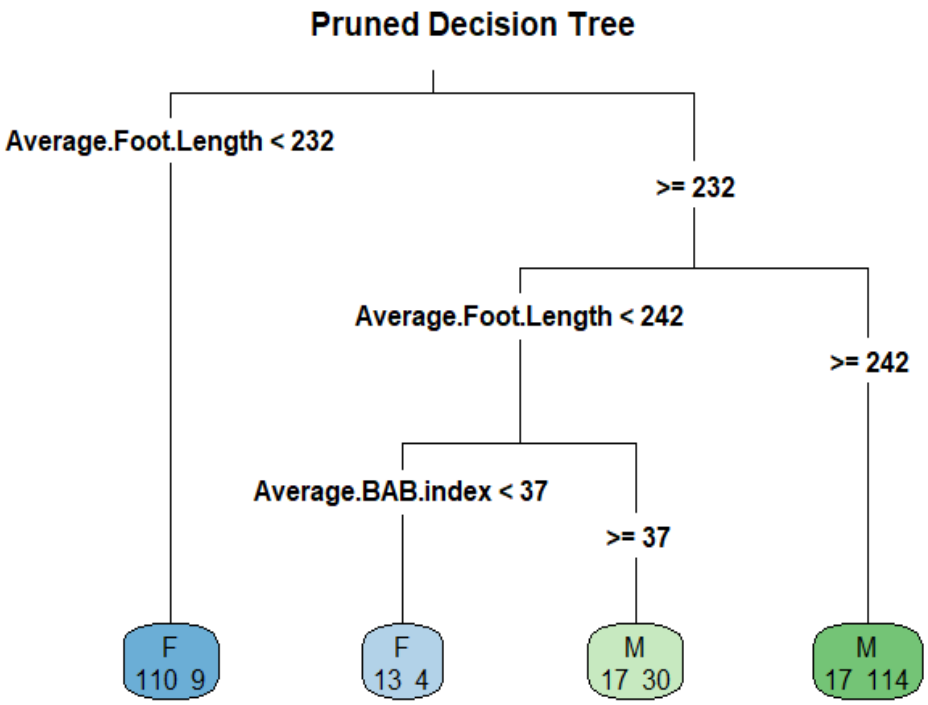
Figure 18 - Visualisation of relative error vs Terminal nodes



In terms of both accuracy and kappa, the model with a cp value of 0.1273885 outperforms the others, indicating a balance between model performance and complexity. As a result, we pruned the tree using 4 terminal nodes and the optimal value to achieve a balance between model complexity and prediction accuracy.

The pruned tree can be found in Figure 19. In comparisons of both trees, we can observe differences in their performance and structural complexity. Using this tree without pruning on the training data, we get an accuracy rate of 86.64%, a sensitivity of 82.80%, and a specificity of 91.08%. Alternatively, the pruned tree has an accuracy of 85.03% and a sensitivity of 78.34%, while maintaining a high specificity of 91.72%. While the pruned trees witnessed a marginal drop in accuracy, they possess several advantages. The simplified structure makes interpretation and generalization easier. Moreover, the primary splits are centred around average foot length and BAB index. It becomes a more practical choice in real-world applications due to its balance between simplicity and performance, making it superior to the unpruned tree despite its marginally high accuracy.

Figure 19 – Visualisation of Pruned Classification tree



Although our 'cp' value gave us good results in our training data, our particular shoe dataset needed more conservative pruning, given the nature of our dataset. After evaluating our pruned tree with cross-validation and other metrics with a cp value of 0.01273885, we chose to set it to 0.02 for the test set. This change was not done randomly but in the pursuit of optimal modal performance of our unseen shoe data. We hoped to find a balance between lowering the tree's complexity to avoid overfitting and keeping its predictive potential. Similar, to what we did in our GNB model we are making the models more conservative to match the archaeologist's predictions of the owner of the shoe.

It is critical to emphasize that the choice reduced to 0.02 was tailored specifically to match the archaeologist's predictions and the test dataset. When applying the model to a new dataset, it becomes important to evaluate, revalidate, and retune it. It emphasizes the necessity of knowing the intricacies of one's data and being adaptable in model creation, ensuring that the parameters and structures chosen are properly aligned with the unique problem domain and the dataset features. The simplicity of this pruned tree shows that the average foot length is the key factor for determining the sex based on the data presented. As a result, it suggests that a foot length less than 232 is more predictive of females, whereas a foot length of 232 or above is more predictive of males. However, this simple

model can only fit our test dataset and can potentially be a hindrance if the data contains more sophisticated patterns that this tree does not represent.

4.7.3 Random Forest –

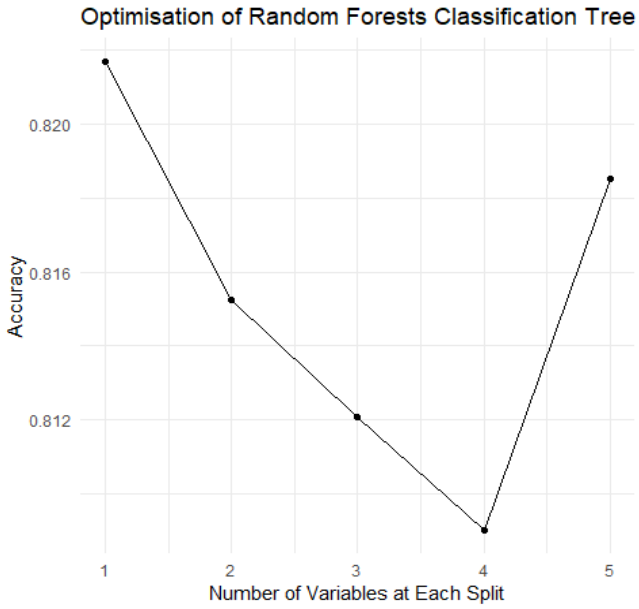
The random forest model was implemented in its default configuration in the initial phase without any modification. During the decision–making process, each spit considered two distinct variables, and the ensemble consisted of 100 trees. In the first attempt, the Out-of-bag error estimate was 21.66%. This model performed 65.82% accurately when compared to the test set (without nan values). Moreover, its accuracy when identifying females referred to sensitivity was 100 %. However, the specificity of its ability to distinguish the males was only 48.08%.

In Figure 20, we can see the visualization of how different configurations can influence the model's performance. Each split contains one to five variables (mtry) on the x-axis, and it displays the accuracy it achieved on the y-axis. First, it achieved an accuracy of 82.17% at 'mtry' of 1, suggesting significant agreement beyond chance. When the mtry increases to 2 it reveals a slight decline in accuracy. As the 'mtry' value increases, the accuracy decreases gradually, with the lowest accuracy of 81% found at 'mtry' of 4. Interestingly, when 'mtry' is 5, we witness a significant improvement in performance registering an accuracy of 81.85%. The plot clearly emphasizes our random forest model's sensitivity to 'mtry' parameter, stressing the tight balance that must be struck for optimal performance.

A decision to use hyperparameter optimization was taken based on the results. Our objective was to tune the 'mtry' value throughout the whole range of possible characteristics.

We used 5-fold cross-validation again to assure the dependability of our evaluation. Our investigation generated intriguing results. Taking only one variable (mtry =1) into account at each split resulted in a progressive climb, achieving an accuracy of 82.17% from 78.34% in the training data. However, as we have seen throughout this analysis the trained model performance differs from our test set. When we ran the trained tuned models, our well-built configuration produced a low accuracy of 54.43%.

Figure 20- Optimisation of Random Forests classification Tree



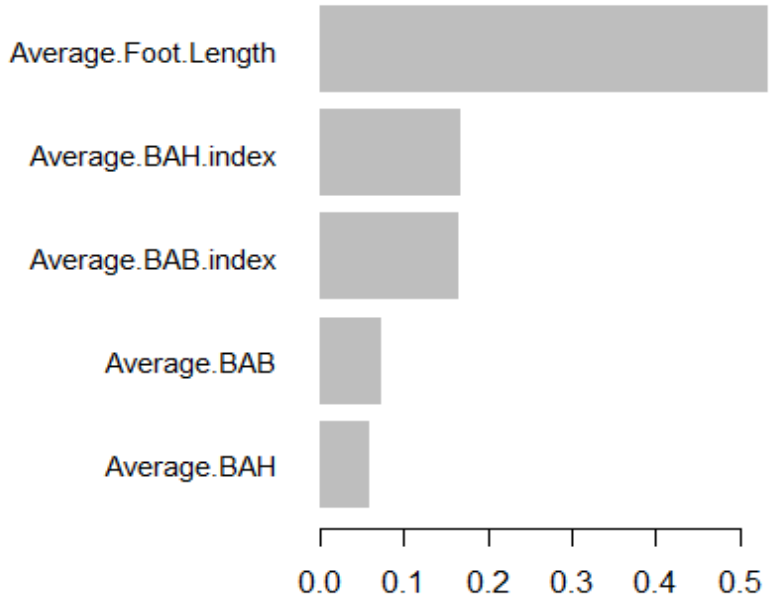
The goal was to improve the performance by tweaking the hyperparameters but our test data performs poorly. As mentioned before, the training set differs from the test set in certain ways. There is a slight distribution shift between these two sets but our primary goal was to make use of the footprint data as much as possible and predict the sex of those shoes.

Despite the tuning efforts, the original random Forest model, which was not adjusted, outperformed the tuned model on the test dataset. This emphasizes the need for practical validation by demonstrating that, while hyperparameter adjustments are beneficial, it does not always ensure a powerful model, especially for our test dataset.

4.7.4 XGBoost

Following our exploration with the random forest model, we employed XGBoost to try to improve our model's performance on unknown data. The switch from Random Forest to XGboost reflects a paradigm change from a parallel ensemble method to a sequential boosting strategy. Before we started training our model, XGBoost can rank features depending on their relevance. In essence, the algorithm assesses how frequently a specific feature appears in the trees as well as its average influence on the model's prediction. These assessments provide scores, which may then be sorted in descending order to produce a ranking of all the characteristics. In Figure 21, the features have been ranked based on their 'F score'.

Figure 21 - Important Variables in XGBoost model



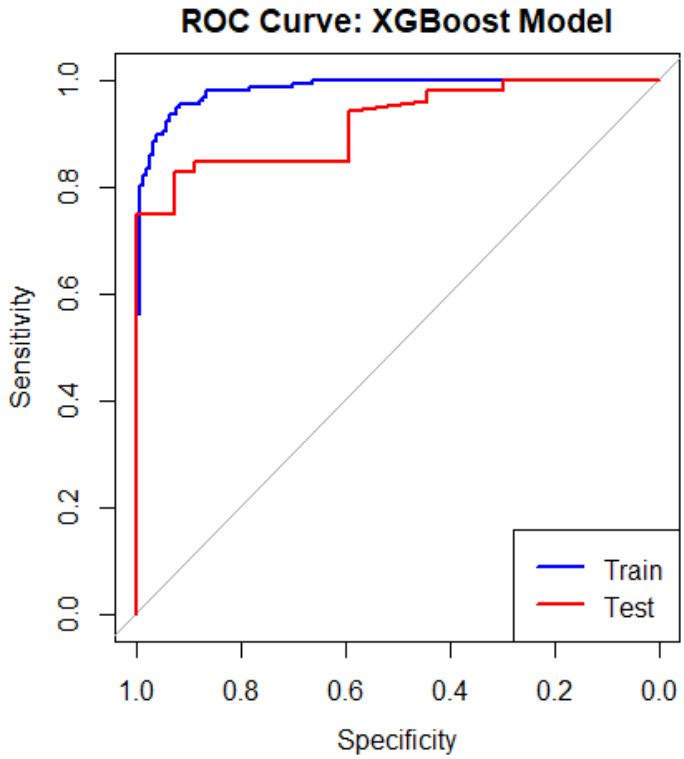
In terms of sex determination, the 'Average Foot Length' stands out prominently, with a significant 44.03% contribution to the model's predictive capability. Averaging BAH index, the next in line, contributes 18.45% of the model's decision-making power, although it's shorter in visual presentation. A significant value of 17.09% is found for the 'average BAB', and 13.41% for the 'average BAB index'. Finally, 'Average BAH' captures 7.01% of the model's gain. Despite their varying lengths, they are each positioned on the plot to represent their individual and collective contributions. Although all of these features are critical to determining sex, their effects differ significantly.

We trained our XGBoost model on ‘balanced_df’ training dataset and used binary logistic objective which is appropriate for our binary classification task. The model was trained over 100 iterations using the error metric as a guiding principle. After training, we evaluated our model on the test set. It achieved an accuracy rate of 58.23% with a sensitivity of 100% and specificity of 36.54%. The model was able to identify all females correctly, but the specificity showed that there was substantial room for improvement in correctly classifying all the males. To guarantee that the model did not overfit the training data and to maximize its prediction performance, we use cross-validation and hyperparameter tweaking. The model was tested on the shoe data after this step. Later, we analyzed the model’s classification performance with ROC curves, which is a fair representation of the model’s true positive and its false positive rate at different thresholds. It indicated an exceptional Area under the curve (AUC) of 91.95%. This implies that our model can distinguish between the two classes. When looking at the ROC data, we can see the balance of sensitivity and specificity at different thresholds. For example, with a threshold of 0.0083, the model has a high specificity of 96.29% which means it

accurately detects 96.29% of the real negatives. However, it is important to note that this does not imply that our model is 96.29% accurate when it comes to the testing data.

The ROC curve is displayed in Figure 22 for a more detailed view.

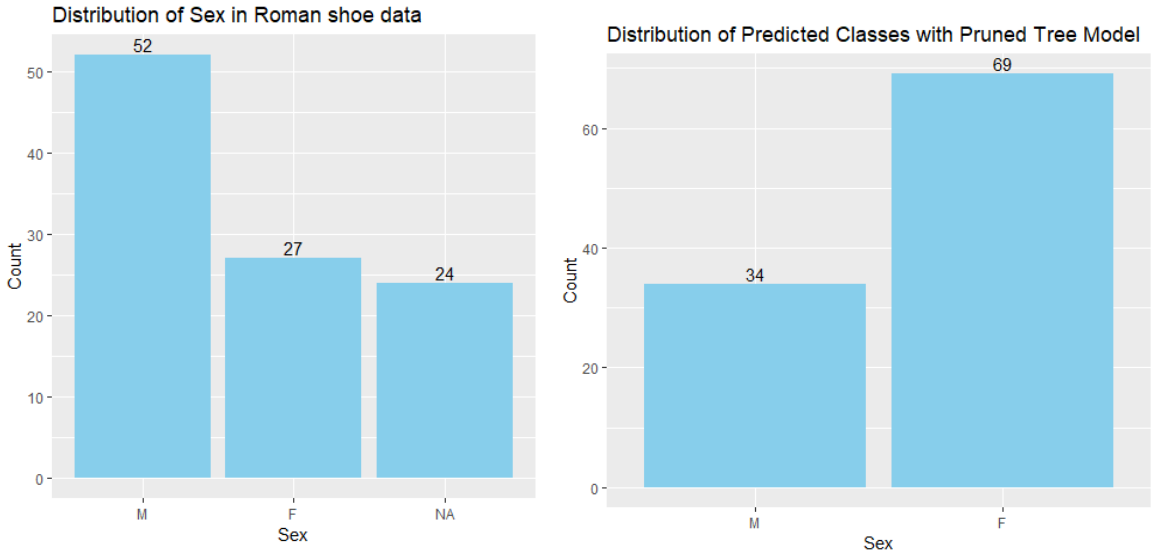
Figure 22 - Receiver operating characteristic curve for XGBoost model



We tuned our XGBoost classifier's hyperparameters rigorously to enhance its performance. It was necessary to fine-tune several parameters of the model's architecture. We optimized the balance between overfitting and model complexity by setting the depth of the trees to a maximum of 4. Following this adjustment, the model performed slightly better than before our tuning on the test set. The accuracy went from 58.23% to 67.09% with a sensitivity (true positive) rate of 100% meaning all positives were correctly identified in terms of balanced accuracy, which takes into account both sensitivity and specificity, the model scored 75%. As a result of tuning hyperparameters, we have seen a notable improvement in our model's ability to classify data points.

As mentioned at the beginning of the classification tree section, we identified the pruned decision tree model to be the top performer in terms of accuracy. The selected models were applied to the shoe dataset with NaN values to see how they perform when they are confronted in a real-world context. By doing this we can test the robustness and predictive power of our trained models by applying them to the dataset without any target variable. Therefore, the objective is not only to fill in missing 'sex' values but also to determine how to apply these models to derive actionable insights.

Figure 23- Best model (Pruned Classification tree) predicting sex of Roman shoes



The shoe dataset included 52 males, 27 females, and 24 entries that were unidentified. On the test set without NaN values, our pruned decision tree model showed an accuracy rate of 76%. Interestingly when we applied the trained model to the dataset with unknown values, 18 of the 52 males were classified as females, as were all 24 entries previously marked as "unknown". While these outcomes may seem unexpected at first, they agree with the model's 76% accuracy rate on the test data. Despite being quite accurate, the model does have limitations, particularly when deviations from the pattern learned during training occur. As a result, even though the predicted distribution differs from the original, the high accuracy rate in the test set provides a very strong reason to trust the models' generalizability.

However, in the majority of the models we trained and tested, "unknown" entries were classified as females. It may be advisable to investigate this pattern further based on its consistency across different models. Either the training data are biased or there is a feature correlated with female characteristics that makes "unknown" entries more likely to be female. To better understand and improve the model's ability to handle such cases, we plan to delve deeper into this pattern.

Investigating the unknown entries -

A thorough investigation of how different models classified "unknown" entries as females revealed some interesting findings. Other than the Bayesian model with thresholds set at 0.1 and 0.4, which classified one data point with an insole length of 240 mm as male, all other models classified these unknowns as female. It is noteworthy that these entries ranged in insole length from 130mm to 190mm. Similarly, the Insole BAB was under 60mm and the mean for the average female BAB in our training data was 85mm.

The insole length range likely corresponds to children, and this limitation of our training data is that we didn't have access to data for children with foot lengths less than 200mm. The youngest subjects in our training set, around 12 years old, had feet that measured over 215mm. In addition, we observed that some adult females had foot lengths around 200 mm, explaining why the model classifies smaller foot lengths as female.

Due to these limitations and observations, it's understandable why our models, which are trained mainly on adult data, identify "unknown" categories as females with shorter insole lengths. This limitation was addressed in our methodology during our dataset formation. Despite aligning with our training data constraints, this finding highlights the need for a larger and more diverse dataset, which includes foot lengths from a range of age groups.

Figure 24 - Histogram of Insole length predicted sex on Roman shoes with unknown entries

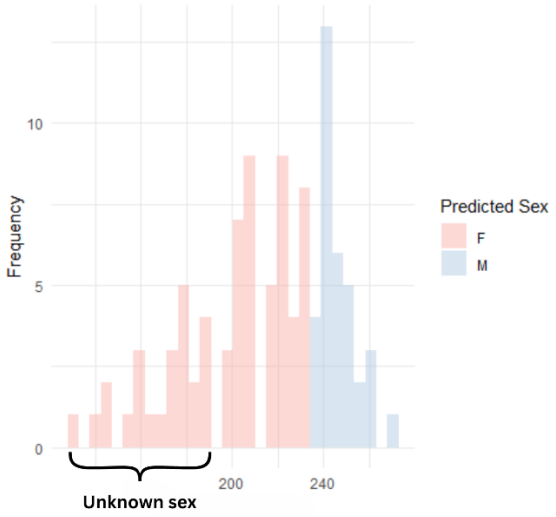
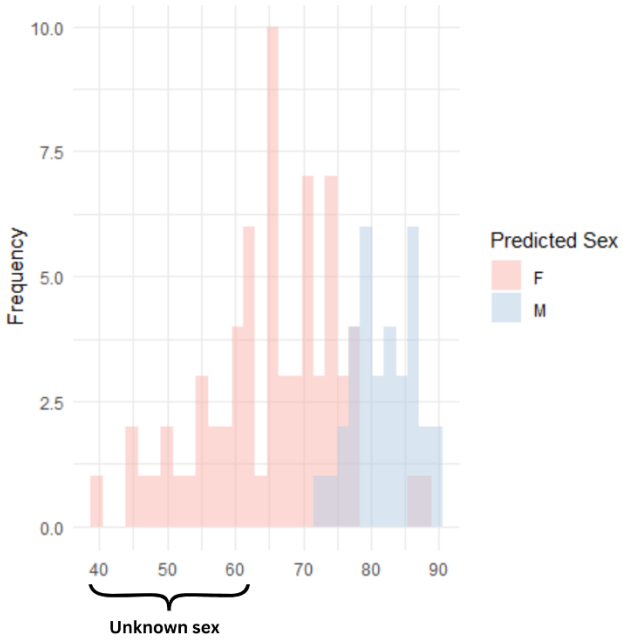


Figure 25- Histogram of Insole BAB predicted sex on Roman Shoes with unknown entries



As a result of this investigation of unknown entries, we are better able to understand the observed tendencies of the machine learning models. We can now look at the strengths, weaknesses, and overall performance of different classification models that we built.

4.8 Model Comparison –

A comparison of the performance of different machine learning models for classifying shoe data based on foot measurements revealed some patterns and differences. Below is a summary table encapsulating key performance metrics for each model –

Table 5- Model comparison Table with type and metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Gaussian NB (0.5 Threshold)	75.95	100.00	63.46	0.78
Gaussian NB (0.6 Threshold)	71.00	100.00	56.00	0.72
Gaussian NB (0.4 Threshold)	87.34	100.00	80.77	0.89
Gaussian NB (0.1 Threshold)	96.20	98.04	96.15	0.97
Classification Tree	58.23	100.00	45.00	0.62
Pruned Tree	75.95	100.00	58.69	0.73
Random Forest	65.82	100.00	50.00	0.67
XGBoost with Tuning	67.09	100.00	50.94	0.68

In our analysis, Gaussian Naive Bayes (NB) with a 0.1 threshold demonstrated the highest accuracy at 96.20%, as well as strong Precision, Recall, and F1-Score metrics. Alternatively, the untuned Classification Tree model showed the lowest accuracy at 58.23%. Although these variations were present, almost all models had a striking similarity. Almost all models had a high precision rate of 100% when it came to identifying females. Our model found it easier to identify females since their foot lengths were shorter. However, we do acknowledge that the results have a slight bias in favour of women. It is believed that this bias stems from the fact that foot length is a very strong signal for identifying females, so it boosts the confidence of the models. However, it is worth noting that there was a slight overlap between males and females on foot sizes. Our predictions diverged slightly from the archaeologists' conclusion because of this factor, when combined with other features in our dataset.

Although the GNB model with a 0.1 threshold displayed the highest accuracy, a model should be able to adapt to unknown data in the real world. This makes the pruned tree and Gaussian NB models with a 0.5 threshold particularly robust. An accuracy of 75.95% was reported by the Pruned Tree model, with a precision of 100% and a recall of 58.69%. In the same way, the Gaussian NB model with a 0.5 threshold had an accuracy of 75.95%, 100% precision, and recall of 63.46%, which was closely

comparable. Both models predict similarly and are equally reliable based on their performance metrics increasing their generalizability with no additional threshold adjustment. In addition, their F1 scores, which measure a model's precision and recall, are relatively high, demonstrating their balanced performance. Their sensitivity and specificity make them ideal for applications involving real-world unlabelled data, especially in scenarios involving sensitivity and specificity.

There were several constraints associated with the dataset available, starting with its size. To allow the models to capture the complexity and nuances in the data features, it was necessary to maximize the use of data for training purposes. In addition, the models were able to generalize better, as evidenced by performance metrics. Although overfitting concerns were raised, the models were tested using a variety of strategies, including oversampling and undersampling. However, these techniques led to poorer accuracy and increased bias, confirming the original split between training and testing. We were concerned about the integrity of our test data. Excessively manipulating or reducing the test data could compromise the integrity of the representation. Despite our awareness of the possibility of overfitting due to the size discrepancy between training and test sets, we felt this approach was the most reasonable in the circumstances. As a result, a balance was struck between the imperatives of empirical validation and real-world applicability while providing robust model training. Across all models, the aggregate metrics indicate an average accuracy of approximately 74.42%, a precision of 99.75%, a recall of 62.5%, and an F1-score of 0.75. The above metrics confirm the general effectiveness of our sex classification models based on the relevant data.

Overall, the machine learning models were able to make predictions based on the available data, which in turn was sourced from archaeological findings. Overall, models without any tuning or adjustments predicted more females. The results do not necessarily imply that the archaeologists' assessments were incorrect but indicate that machine learning algorithms trained on the dataset used were more likely to predict a higher number of females. As a result, our training data may indicate the models' ability to generalize well.

5. Discussion

Our study's primary goal was to utilize modern foot measurements from European and Ghanaian populations to build predictive models that would estimate Sex of Roman Shoes. Based on our initial hypothesis, this diverse foot morphology enhanced the learning capabilities of all our models. We were able to demonstrate through statistical analysis that Ghanaians and European foot morphologies are statistically different which justifies their inclusion would increase the diversity in the training data. Moreover, the paired T-Test concluded that there was not much difference between the left and right foot dimensions for both males and females. This simplified our training data because we averaged the measurements of the left foot and right foot. The preparation of the training data for males and females by addressing class imbalance, outliers and distribution normality were aligned with the minimum and maximum measurements of the test dataset. After observing both our training set and testing set distribution, it is evident that there is a marked difference in foot length between males and females. The average foot length for males registered around 248mm and 227mm for females. This difference in foot length highlights as an important indicator for sex but BAH and BAB play an important role in picking up patterns. A little overlap was intentionally retained in our training set to ensure that our model was not too biased and could handle real-world data. This analysis concludes that while foot length can be a primary indicator, it cannot be the sole determinant. Including additional data and measurements is critical for increasing the reliability of predictive models attempting to determine sex based on foot parameters.

We also evaluated a variety of machine learning models, including the Gaussian naïve Bayes model and different classification trees and found noteworthy results, among others in accordance with our second objective in this study. The models that were constructed predicted more women compared to archaeologists' predictions using traditional methods. The shoe dataset without the unknown values contained predictions of 52 males and 27 females, while most of our models predicted around 29-33 males and 46-50 females. Compared to the initial archaeological predictions, our models predicted a 26.58% lower proportion of males and 26.58% higher proportion of females. This considerable change in sex distribution draws attention to how machine-learning models and conventional archaeology approaches differ in their underlying assumptions and methodologies. We also noticed that when we applied the trained model to the whole dataset, it predicted all the entries with foot lengths below 200mm as females which can possibly be children. However, there was an average accuracy of 74% that was maintained across all the models which indicated a good performance. The inclusion of multiple foot measurements can be responsible for the difference between the predictions of our models and the archaeologist's predictions. Foot length has been typically the core factor to determine sex in traditional archaeological methods like Van–Driel Murray's model. On the other hand, our explanatory Data analysis revealed substantial relationships with foot length, Breadth at

Ball (BAB), and Breadth at Heel (BAH). Through the use of all these variables, our models were able to identify different patterns that were inaccessible using traditional univariate methods. Although, it is important to note that Foot length is still the core variable. Therefore, while Van-Driel Murray's approach to assigning sex using foot length is fundamentally sound, our analysis suggests that including other variables provides a more comprehensive understanding when it comes to differentiating male and female foot morphologies.

To predict more males in the shoe dataset and align it closely to the archaeologist's predictions, we had to reduce the threshold in the GNB model. This made our models more conservative in predicting more males, and it also raised concern about the dataset's inherent imbalances and biases. If the model needs to be manipulated to match more closely the with the initial predictions, it triggers a revaluation not only of the data but also of the traditional approaches of archaeologists for interpretation.

Based on our research results, all our models consistently revealed a higher number of women than men in predicting the sex for shoes. These results add weight and support the views of Van-Driel Murray and Greene, who have proposed and published papers on the role and presence of women and argue against the traditional assumption that the fort's demographic was mostly male soldiers and some families. Their findings pointed to a larger social makeup, with a significant presence of women who may have been more than just wives of soldiers and played varied social and economic roles inside and around the fort. However, it is important to acknowledge the limitation which was the sample size of 103 shoes. Despite this, the consistency of our results in all the models provides a solid foundation for the potential of employing shoes as trustworthy markers for predicting sex and understanding the demographic in Vindolanda Fort by using shoes as a proxy for machine learning models.

5.1 Limitations

One major disadvantage in our analysis is the limited number of available foot measurement samples. There was insufficient data for children under the age of 12 and our model was biased in predicting the sex of children's shoes. This specific demographic may not effectively represent the entire population, limiting the generalizability of our findings. Data augmentation was initially considered to expand the size of our training set, but given the nature of our research and dataset, it is both historical and anthropological, hence, introducing new data points artificially was an option. Each sample is valuable due to the cultural and biological peculiarities associated with foot features like size and shape, and generating artificial samples has the potential to introduce a level of inaccuracy or bias that would affect the study's validity. The historical context of Romans and anthropological variations in foot shape and size across cultures and time were also reasons to not go with data augmentation. As a result, despite its limited size and scope, we proceeded with our analysis with the initial dataset,

stressing the need for more samples for future studies. Similarly, the shoe data sample size was small and contained numerous missing values of the insole measurements limiting the dataset. A reduced sample size in the test set may introduce statistical irregularities and is more prone to overfitting.

We validated our models against the sex assigned by archaeologists, but if these assignments are incorrect or biased, this would affect the validation metrics of our analysis. While our study included a variety of foot measurements as indicators, other unmeasured variables like the circumference of the foot around the arch, and heel diagonal may have contributed to additional context or accuracy of our classification models. Moreover, the models are unable to explain why some of the features BAH or BAB are more responsible for predicting sex unlike, traditional research approaches that may have a biological argument or social argument.

One of the most notable limitations of this study is the time period difference between modern Europeans and Ghanaians that was used for training the models and the Roman shoes. While the foot structure may exhibit some universal characteristics, it is important to acknowledge that cultural footwear, lifestyle, nutrition and other environmental factors could influence the shape and size of the foot. These elements have most likely evolved or diversified over centuries. Thus, the feet that wore these Roman shoes belonged to a different environment than the modern population which would have tampered with their foot/ bone structure.

5.2 Conclusion and Final Thoughts –

In conclusion, with the power of machine learning and human expertise, we were able to estimate the sex on the Roman Shoes found in Vindolanda. There was a slight stature difference between the two populations which hints at the variability and potential factors that have influenced us over time. The study did possess some implications. It invites archaeologists to reexamine long-held assumptions by harnessing the power of data analytics, and for data science professionals, it emphasizes the importance of subject expertise. The combination of both would be a huge contributing factor in the field of archaeology.

However, we were able to predict the sex of 79 shoes, excluding the unknown data which are classified as possible children in our study. This discrepancy highlights the challenges in obtaining firm conclusions from predictive algorithms. Regardless, we were able to provide some useful insights about the possible sex distribution in Vindolanda Fort, indicating a strong female presence.

6. References

10.7 - Detecting multicollinearity using variance inflation Factors | STAT 462. (n.d.). https://online.stat.psu.edu/stat462/node/180/

Abledu, J. K., Abledu, G. K., Offei, E. B., & Antwi, E. M. (2015). Determination of Sex from Footprint Dimensions in a Ghanaian Population. *PLOS ONE*, *10*(10), e0139891. https://doi.org/10.1371/journal.pone.0139891

Alberti, M. (2018). The construction, use, and discard of Female Identities: Interpreting spindle whorls at Vindolanda and Corbridge. *Theoretical Roman Archaeology Journal*, *1*(1), 2. https://doi.org/10.16995/traj.241

Allason-Jones, L. (1995). 'Sexing' small finds. *Theoretical Roman Archaeology Journal*, *0*(1992), 22. https://doi.org/10.16995/trac1992_22_32

Allason-Jones, L. (2012). Women in Roman Britain. *A Companion to Women in the Ancient World (*, 467–477. https://doi.org/10.1002/9781444355024.ch34

Atamtürk, D. (2010). Estimation of Sex from the Dimensions of Foot, Footprints, and Shoe. *Anthropologischer Anzeiger*, *68*(1), 21–29. https://doi.org/10.1127/0003-5548/2010/0026

Attia, M. H., & Abulnoor, B. a. E. (2019). Sex estimation of femur using simulated metapopulation database: A preliminary investigation. *Forensic Science International: Reports*, *1*, 100009. https://doi.org/10.1016/j.fsir.2019.100009

Attia, M. H., Kholief, M., Zaghloul, N., Kružić, I., Anđelinović, Š., Bašić, Ž., & Jerković, I. (2022). Efficiency of the adjusted binary Classification (ABC) approach in osteometric sex Estimation: A comparative study of different linear machine learning algorithms and training sample sizes. *Biology*, *11*(6), 917. https://doi.org/10.3390/biology11060917

Awais, M., Naeem, F., Rasool, N., & Mahmood, S. (2018). Identification of sex from footprint dimensions using machine learning: a study on population of Punjab in Pakistan. *Egyptian Journal of Forensic Sciences*. https://doi.org/10.1186/s41935-018-0106-2

Bickler, S. H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, *9*(2), 186–191. https://doi.org/10.1017/aap.2021.6

Buck, T., Greene, E. M., Meyer, A., Barlow, V., & Graham, E. (2019). The body in the ditch: alternative funerary practices on the northern frontier of the Roman Empire? *Britannia*, *50*, 203–224. https://doi.org/10.1017/s0068113x1900014x

Ca, O., Lb, B., & Ikpa, J. O. (2020). Predictive Models for Sex and Stature Estimation using Foot Anthropometric Dimensions among Indigenes of. . . *ResearchGate*. https://www.researchgate.net/publication/346084435_Predictive_Models_for_Sex_and_Stature_Estimation_using_Foot_Anthropometric_Dimensions_among_Indigenes_of_Cross_River_State

Case, D. T., & Ross, A. H. (2007). Sex Determination from Hand and Foot Bone Lengths. *Journal of Forensic Sciences*, *52*(2), 264–270. https://doi.org/10.1111/j.1556-4029.2006.00365.x

Coelho, J. D., & Curate, F. (2019). CADOES: An interactive machine-learning approach for sex estimation with the pelvis. *Forensic Science International*, *302*, 109873. https://doi.org/10.1016/j.forsciint.2019.109873

Curate, F., Coelho, J. D., Gonçalves, D., Coelho, C., Ferreira, M. T., Navega, D., & Cunha, E. (2016). A method for sex estimation using the proximal femur. *Forensic Science International*, *266*, 579.e1-579.e7. https://doi.org/10.1016/j.forsciint.2016.06.011

De Boer, H. H., Obertová, Z., Cunha, E., Adalian, P., Baccino, E., Fracasso, T., Kranioti, E. F., Lefèvre, P., Lynnerup, N., Petaros, A., Ross, A. H., Steyn, M., & Cattaneo, C. (2020). Strengthening the role of forensic anthropology in personal identification: Position statement by the Board of the Forensic Anthropology Society of Europe (FASE). *Forensic Science International*, *315*, 110456. https://doi.org/10.1016/j.forsciint.2020.110456

Del Bove, A., & Veneziano, A. (2022). A Generalised Neural Network Model to Estimate Sex from Cranial Metric Traits: A Robust Training and Testing Approach. *Applied Sciences*, *12*(18), 9285. https://doi.org/10.3390/app12189285

Delgado, Y., Price, B. S., Speaker, P. J., & Stoiloff, S. (2021). Forensic intelligence: Data analytics as the bridge between forensic science and investigation. *Forensic Science International: Synergy*, *3*, 100162. https://doi.org/10.1016/j.fsisyn.2021.100162

Garvey, R. (2018). Current and potential roles of archaeology in the development of cultural evolutionary theory. *Philosophical Transactions of the Royal Society B, 373*(1743), 20170057. https://doi.org/10.1098/rstb.2017.0057

Greene, E. M. (2013). Female Networks in Military Communities in the Roman West: A View from the Vindolanda Tablets. In *BRILL eBooks* (pp. 369–390). https://doi.org/10.1163/9789004255951_020

Greene, E. M. (2015). *Conubium cum uxoribus*: wives and children in the Roman military diplomas. *Journal of Roman Archaeology*, *28*, 125–159. https://doi.org/10.1017/s1047759415002433

Harris, W. V. (1980). Towards a study of the Roman slave trade. *Memoirs of the American Academy in Rome*, *36*, 117. https://doi.org/10.2307/4238700

Hoey, C., Wang, A., Raymond, R. J., Ulagenthian, A., & Kryger, K. O. (2022). Foot morphological variations between different ethnicities and sex: a systematic review. *Footwear Science*, *15*(1), 55–71. https://doi.org/10.1080/19424280.2022.2148294

Hörr, C., Lindinger, E., & Brunnett, G. (2014). Machine learning based typology development in archaeology. *Journal on Computing and Cultural Heritage*, *7*(1), 1–23. https://doi.org/10.1145/2533988

Knecht, S., Santos, F., Ardagna, Y., Alunni, V., Adalian, P., & Nogueira, L. (2023). Sex estimation from long bones: a machine learning approach. *International Journal of Legal Medicine*. https://doi.org/10.1007/s00414-023-03072-4

Krishan, K., Chatterjee, P. M., Kanchan, T., Kaur, S., Baryah, N., & Singh, R. (2016). A review of sex estimation techniques during examination of skeletal remains in forensic anthropology casework. *Forensic Science International*, *261*, 165.e1-165.e8. https://doi.org/10.1016/j.forsciint.2016.02.007

Luo, G., Houston, V. L., Mussman, M., Garbarini, M., Beattie, A. C., & Thongpop, C. (2009). Comparison of male and female foot shape. *Journal of the American Podiatric Medical Association*, *99*(5), 383–390. https://doi.org/10.7547/0990383

Morrison, G. S., Weber, P., Basu, N., Puch-Solis, R., & Randolph-Quinney, P. (2021). Calculation of likelihood ratios for inference of biological sex from human skeletal remains. *Forensic Science International: Synergy*, *3*, 100202. https://doi.org/10.1016/j.fsisyn.2021.100202

Nagano, K., Okuyama, R., Taniguchi, N., & Yoshida, T. (2018). Gender difference in factors affecting the medial longitudinal arch height of the foot in healthy young adults. *Journal of Physical Therapy Science*, *30*(5), 675–679. https://doi.org/10.1589/jpts.30.675

Nikita, E., & Nikitas, P. (2019). Sex estimation: a comparison of techniques based on binary logistic, probit and cumulative probit regression, linear and quadratic discriminant analysis, neural networks, and naïve Bayes classification using ordinal variables. *International Journal of Legal Medicine*, *134*(3), 1213–1225. https://doi.org/10.1007/s00414-019-02148-4

Orr, C., Williams, R., Halldórsdóttir, H., Birley, A., Greene, E. M., Nelson, A., Ralebitso-Senior, T. K., & Taylor, G. (2021). Unique chemical parameters and microbial activity lead to increased archaeological preservation at the Roman frontier site of Vindolanda, UK. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-94853-7

Oura, P., Korpinen, N., Machnicki, A. L., & Junno, J. (2023). Deep learning in sex estimation from a peripheral quantitative computed tomography scan of the fourth lumbar vertebra—a proof-of-concept study. *Forensic Science Medicine and Pathology*. https://doi.org/10.1007/s12024-023-00586-6

Quade, L., & Gowland, R. (2021). Height and health in Roman and Post-Roman Gaul, a life course approach. *International Journal of Paleopathology*, *35*, 49–60. https://doi.org/10.1016/j.ijpp.2021.10.001

Redfern, R., DeWitte, S. N., Pearce, J., Hamlin, C., & Dinwiddy, K. E. (2015). Urban-rural differences in Roman Dorset, England: A bioarchaeological perspective on Roman settlements. *American Journal of Physical Anthropology*, *157*(1), 107–120. https://doi.org/10.1002/ajpa.22693

Scipy.Stats.Probplot — SciPY v1.11.2 Manual. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html

Sex estimation of the human skeleton. (2020). In *Elsevier eBooks*. https://doi.org/10.1016/c2017-0-03550-4

Stanley, L. (2020). An Empire of Letters: the Vindolanda Tablets, Epistolarity, and Roman Governance. *The Journal of Epistolary Studies*, *2*(1). https://doi.org/10.51734/jes.v2i1.25

Tomassoni, D., Traini, E., & Amenta, F. (2014). Gender and age related differences in foot morphology. *Maturitas*, *79*(4), 421–427. https://doi.org/10.1016/j.maturitas.2014.07.019

Toneva, D., Nikolova, S., Agre, G., Zlatareva, D., Hadjidekov, V., & Lazarov, N. (2020). Data mining for sex estimation based on cranial measurements. *Forensic Science International*, *315*, 110441. https://doi.org/10.1016/j.forsciint.2020.110441

Toy, S., Secgin, Y., Oner, Z., Turan, M. K., Oner, S., & Senol, D. (2022). A study on sex estimation by using machine learning algorithms with parameters obtained from computerized tomography images of the cranium. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-07415-w

Van Driel-Murray, C. (1995). Gender in question. *Theoretical Roman Archaeology Journal*, *0*(1992), 3. https://doi.org/10.16995/trac1992_3_21

Van Driel-Murray, C. (2001). Vindolanda and the dating of Roman footwear. *Britannia*, *32*, 185. https://doi.org/10.2307/526955

Vernon, W., Reel, S., & Howsam, N. (2020). Examination and Interpretation of Bare Footprints in Forensic Investigations *Research and Reports in Forensic Medical Science, Volume 10*, 1–14. https://doi.org/10.2147/rrfms.s241264

7. Appendix

7.1 Python script

Python code and functions for Gaussian Naïve Bayes model and Explanatory Data Analysis. The following is the python script for the Data cleaning, Data alignment, Normality, GNB model and EDA:

https://github.com/code-hobbit/Python_code-for-Roman-shoes-/blob/510cf4e5bbb119387109589f168d482bd9f1c5fb/Python%20code%20GNB%20model%20and%20EDA

7.2 R Script

The following is the R code for classification trees:

https://github.com/code-hobbit/Python_code-for-Roman-shoes-/blob/dcf2349e0b4aac90916707fe677d813f46997aee/R%20studio%20code

7.3 Descriptive statistics of Male and Female after Data set alignment

	Male Average foot Length:	Female Average foot Length:
count	157.000000	157.000000
mean	248.238854	227.259873
std	9.820926	12.265168
min	218.500000	204.500000
25%	241.000000	218.000000
50%	249.500000	226.500000
75%	257.000000	234.500000
max	266.500000	262.000000
	Male Average BAB Stats:	Female Average BAB Stats:
count	157.000000	157.000000
mean	94.668790	85.549682
std	5.500453	6.170168
min	80.000000	71.000000
25%	90.500000	81.000000
50%	94.500000	86.000000
75%	98.000000	89.500000
max	107.000000	101.600000
	Male Average BAH Stats:	Female Average BAH Stats:
count	157.000000	157.000000
mean	52.700637	47.145223
std	4.974892	5.286078
min	42.500000	35.500000
25%	49.000000	43.000000
50%	52.500000	47.000000
75%	56.000000	50.500000
max	65.500000	61.500000

7.4 Table of Figures

7.5 Table of Tables