# "Hominid Chronology and Classification: Machine Learning Applications in Paleoanthropology"

•••

# Challenges in Classification and Dating in Paleoanthropology

One of the biggest challenges in paleoanthropology is identifying and classifying discovered fossils. Hominids exhibit significant variation within the same species, which can lead to confusion about whether certain fossils represent a new species or simply intraspecific diversity.

Another major challenge is fossil dating. Traditional techniques, such as radiocarbon dating or isotope analysis of surrounding rocks, are often expensive, invasive, or limited by the availability of suitable materials. In this context, data science techniques that correlate physical traits with known time periods can provide a non-invasive and cost-effective tool to complement these traditional methods.

# The Project

This project aims to develop predictive models that classify hominid fossils in terms of their Genus and Species and estimate their geological age using variables related to their physical characteristics. This serves as a preliminary analytical tool to complement traditional paleoanthropological research techniques.

Objectives:

- Taxonomic Classification of Fossils:
  Design and implement supervised classification models that utilize physical traits—such as cranial size, bone length, and endocranial capacity—to accurately predict the Genus and Species of fossils.
- Chronological Estimation:
  Develop regression models to estimate the geological age (approximate time period) of hominid fossils based on the same physical characteristics or derived variables.

# The Data

For this project, we used a hominid dataset sourced from Kaggle, which provides a comprehensive collection of data related to the physical, contextual, and chronological characteristics of fossil specimens. This dataset was chosen due to its richness in relevant information for the project's objectives, including taxonomic classification (Genus_Species) and chronological estimation (Time).

The dataset contains 12,000 records and 27 columns, offering a robust framework for analysis and modeling. Each row represents a fossil specimen, while the columns reflect various associated attributes.

Dataset used in this study:
🔗 Kaggle - Biological Data of Human Ancestors

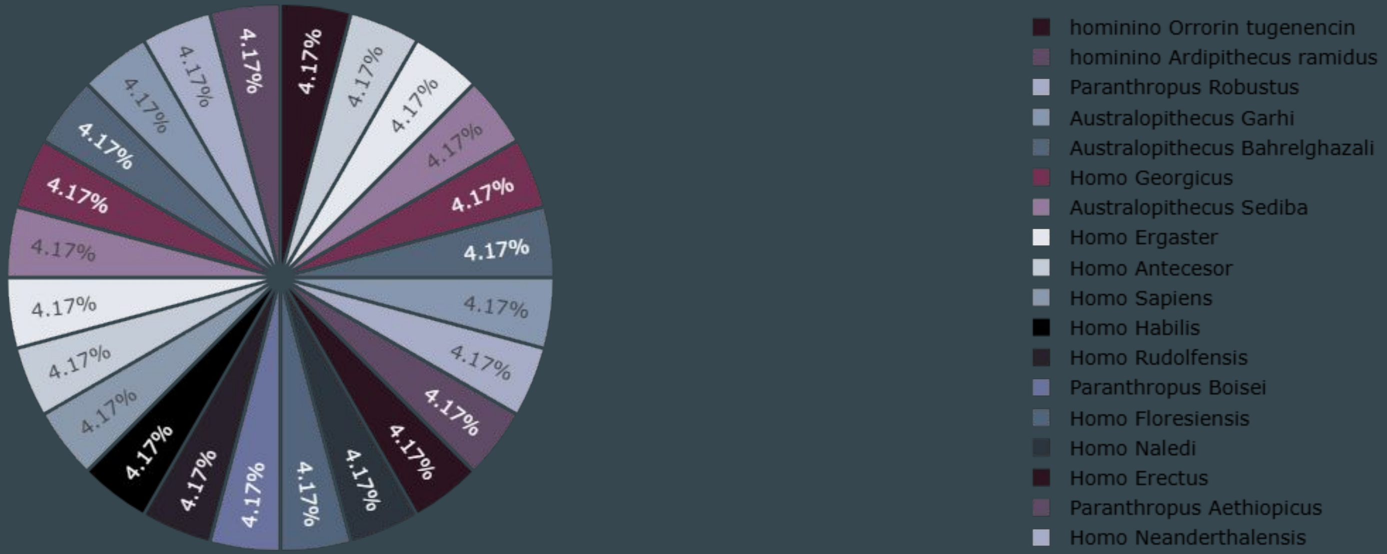
# Data Cleaning

The data cleaning process for this project did not require imputing missing values or handling null entries, as the dataset contained no incomplete records. However, several key transformations were performed to enhance data consistency and usability:

- Column Name Standardization:
  Column names were modified to ensure consistency and clarity, using a uniform and descriptive naming convention. This improved the readability and handling of variables during analysis.
- Time Variable Transformation:
  The time variable was standardized to always be expressed as a numerical value representing the number of millions of years before the present.
- Previously, time was recorded in various formats, including ranges and textual descriptions, which complicated the analysis.
- Cranial Capacity Standardization:
  The cranial capacity variable was normalized to ensure all values were expressed in cubic centimeters ($cm^3$).

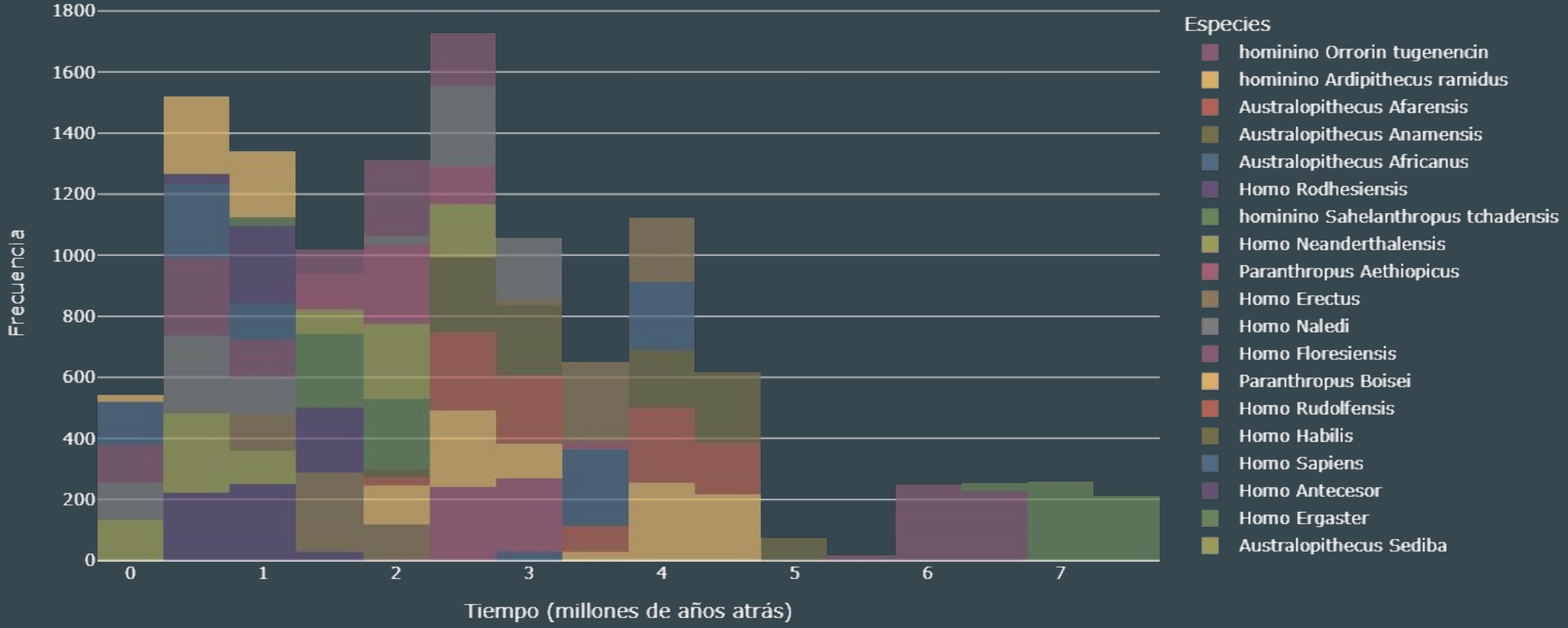
# EDA (Exploratory Data Analysis)

This balance ensures that no single species dominates the analysis, allowing the physical and temporal characteristics of each group to be evaluated equitably.

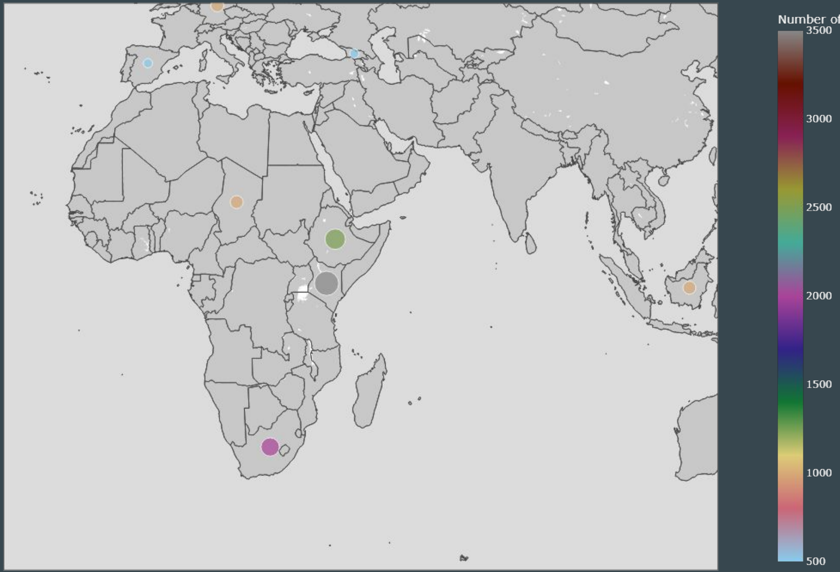


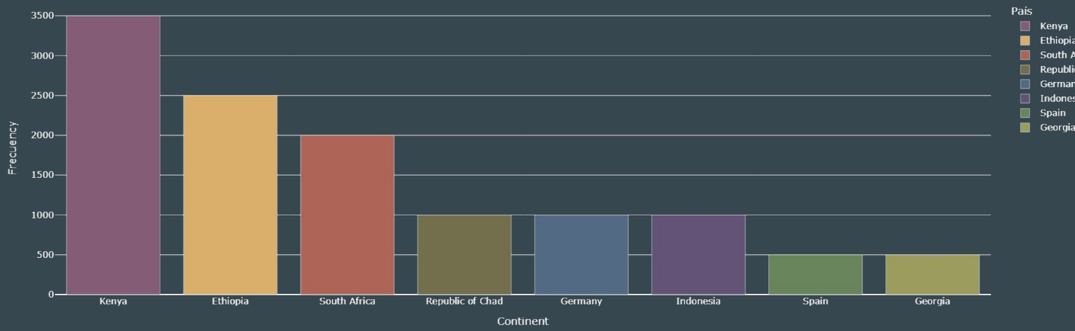
# Species Over Time

Frequency of fossil records in each time interval.

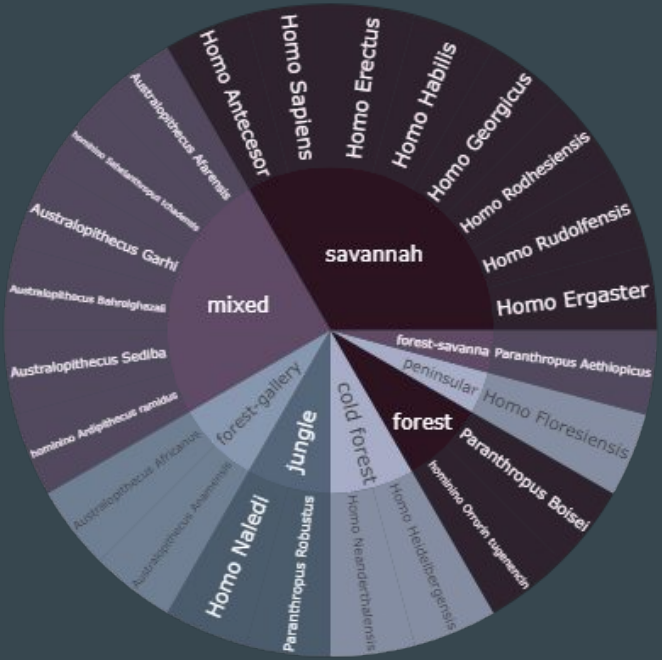


# Number of Fossils by Country





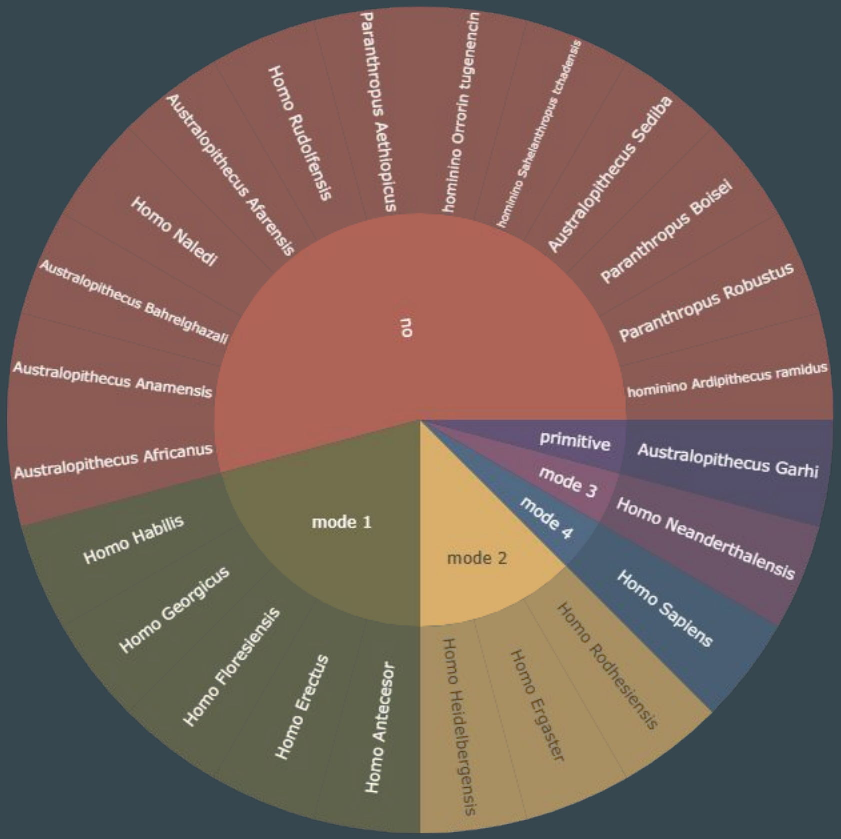
# Habitat by Species



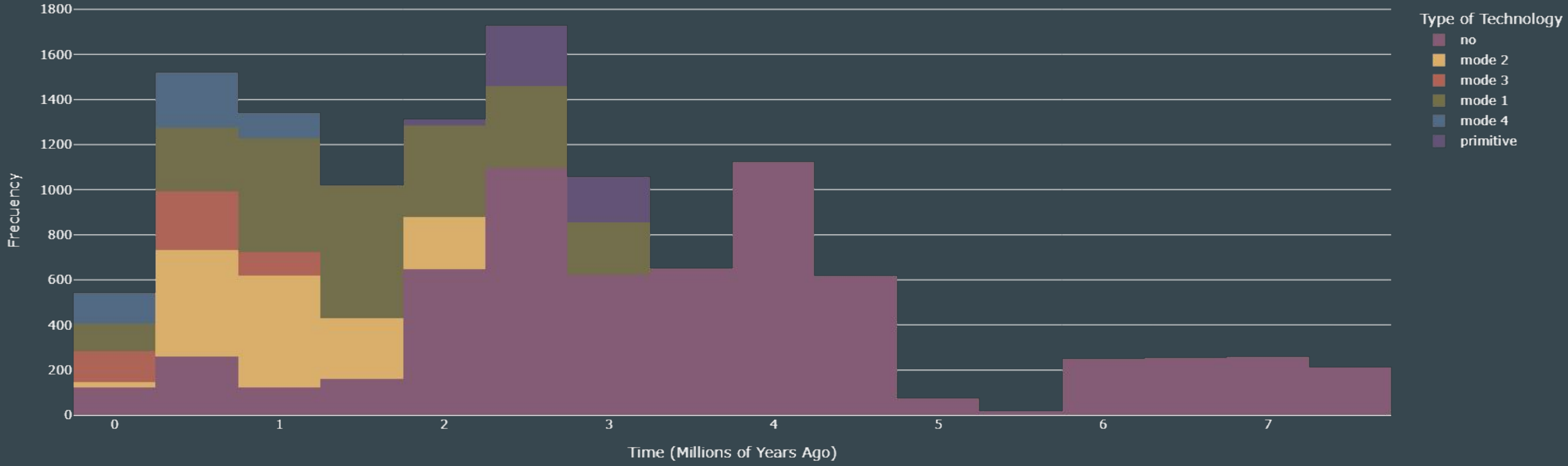
There is a close relationship between hominid species diversity and their ecological environments. The predominance of **savanna habitats** highlights their role as a key evolutionary cradle, closely associated with species such as *Homo erectus, Homo habilis,* and *Homo ergaster*.

Meanwhile, more specific habitats reflect how different species adapted to varying environmental pressures. This provides a comprehensive evolutionary perspective on how hominids interacted with and thrived in their surroundings.

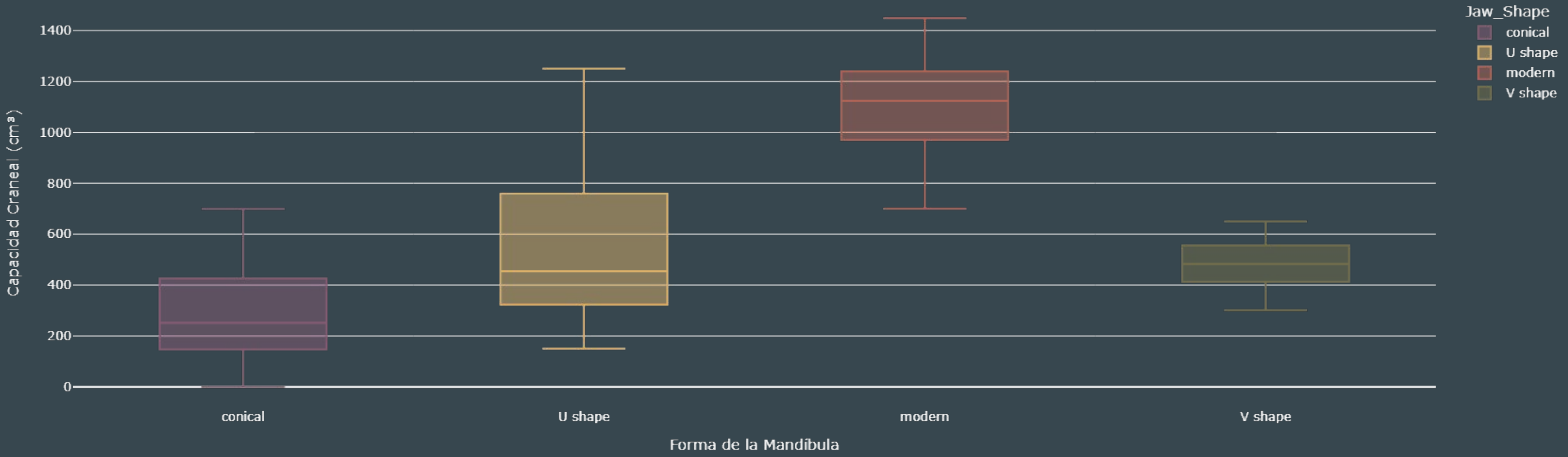
## Hierarchical Distribution of Tecno_Type and Genus & Species



# Type of Technology Over Time

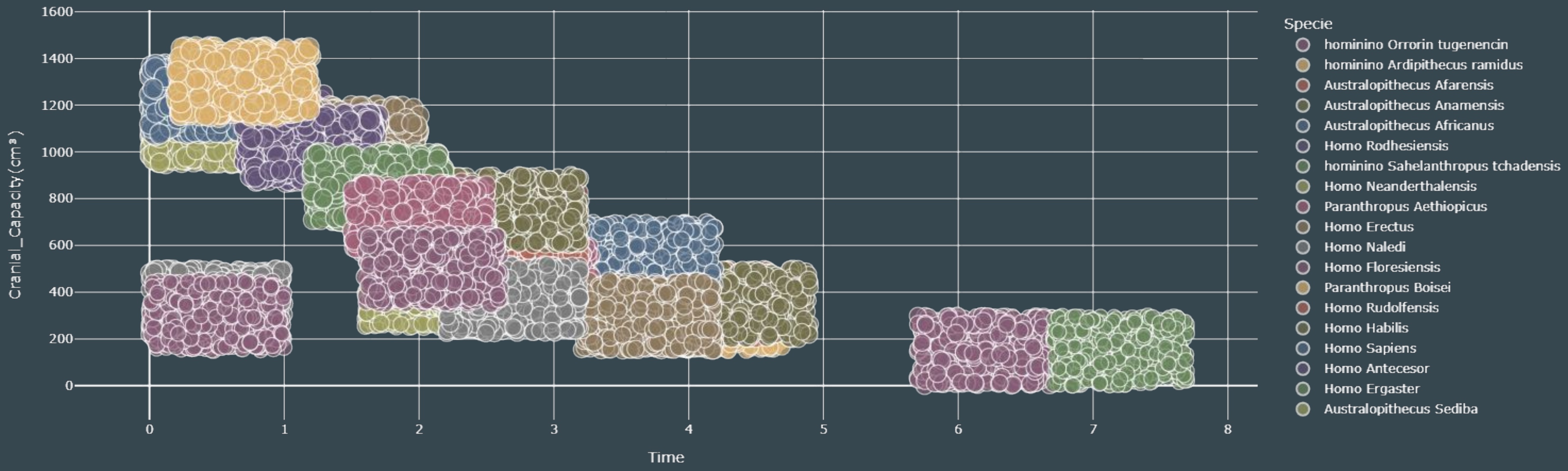


# Relationship Between Cranial Capacity and Jaw Shape

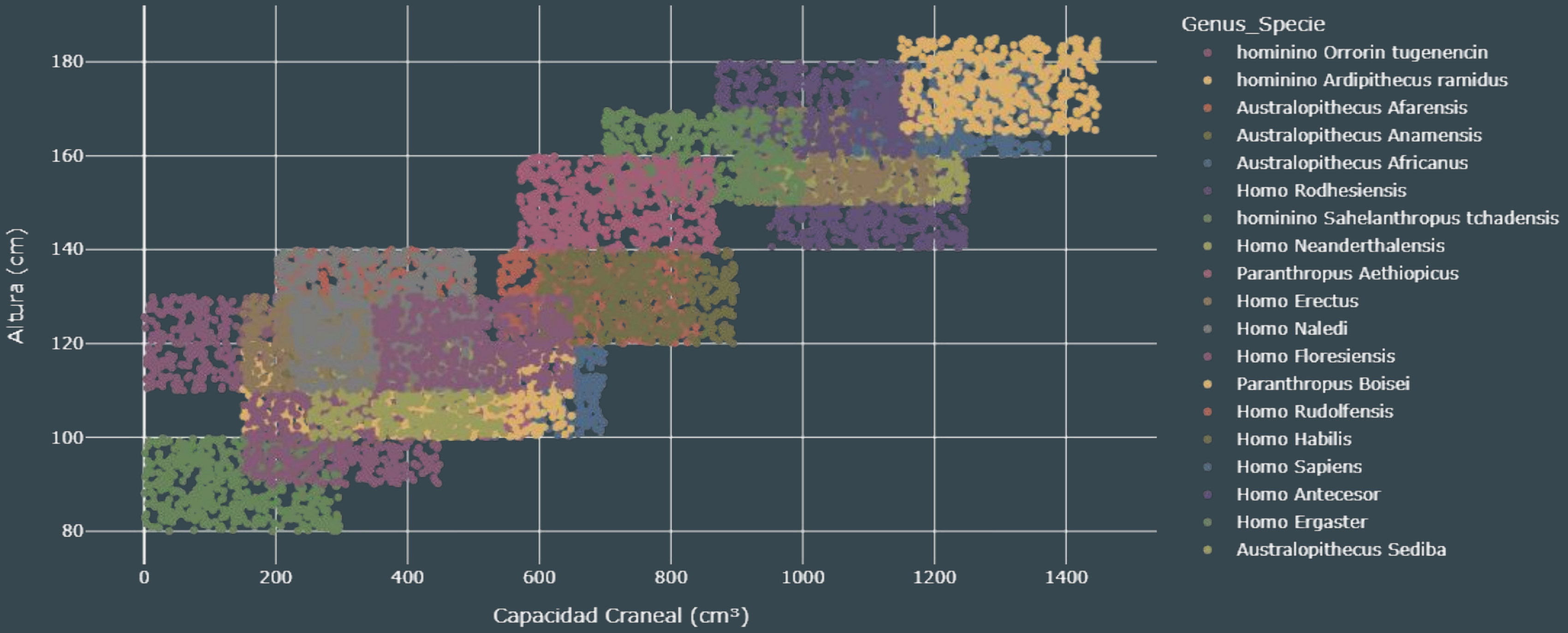


# Relationship Between Geological Time (in Millions of Years) and Cranial Capacity

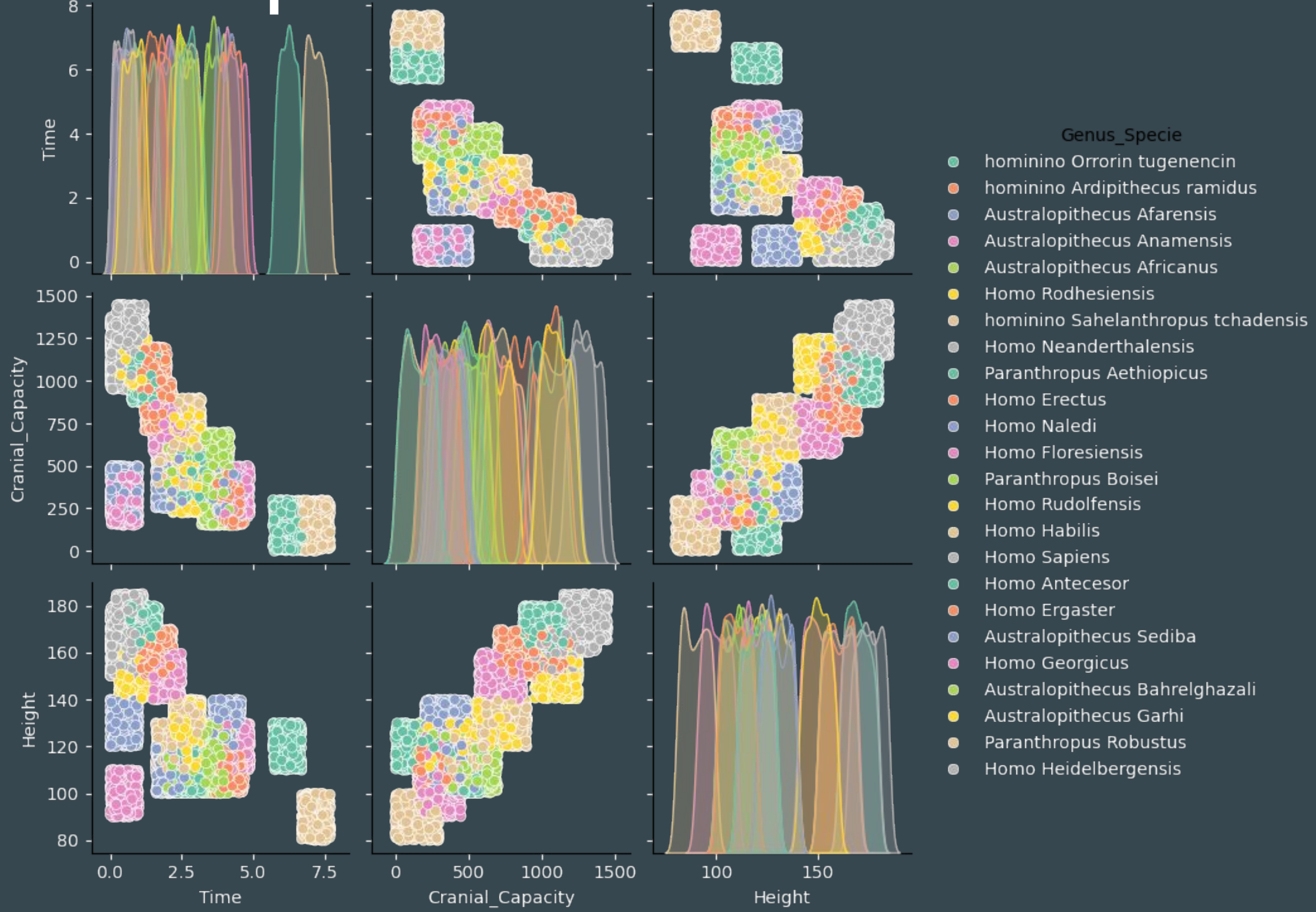
Evolutionary Trend: A progressive increase in cranial capacity is observed as time decreases (more recent species).



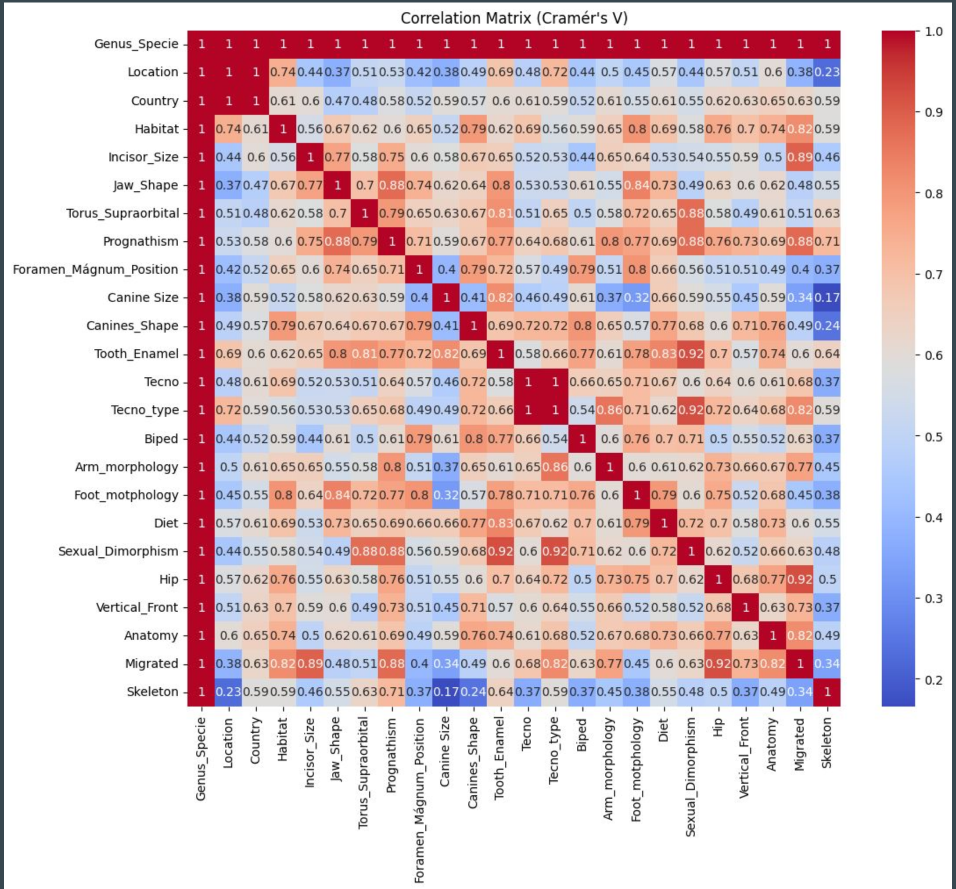
# Correlation Between Cranial Capacity and Height



# Relationship Between Numerical Variables



# Correlation of Categorical Variables (Cramér's V)

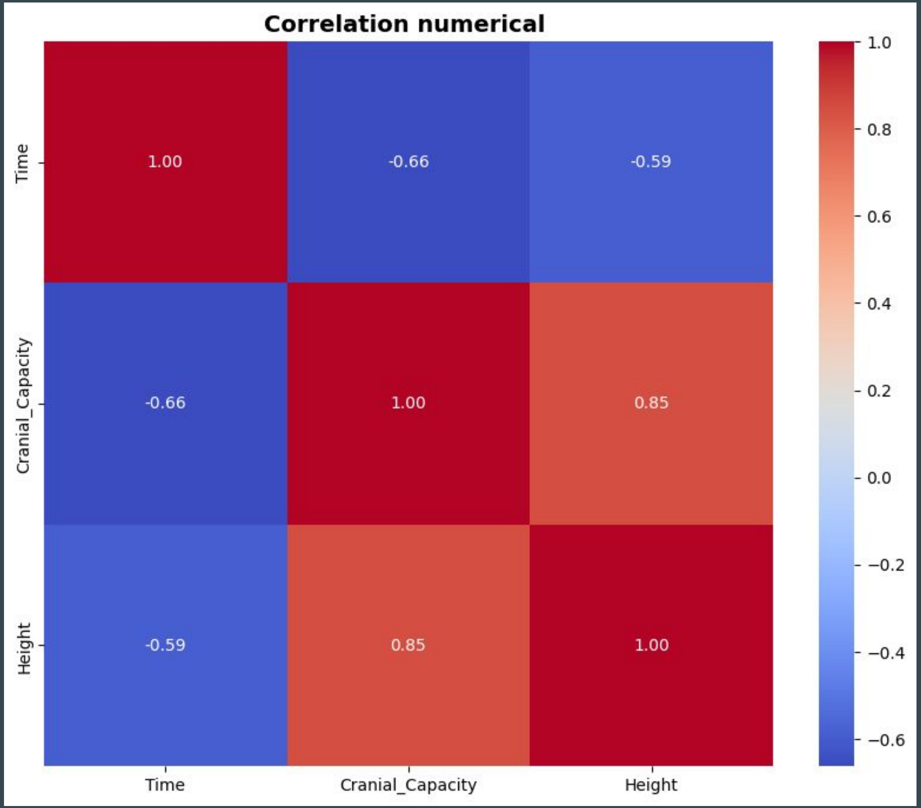


Correlation Matrix (Cramér's V)

| | Genus_Specie | Location | Country | Habitat | Incisor_Size | Jaw_Shape | Torus_Supraorbital | Prognathism | Foramen_Mágnum_Position | Canine Size | Canines_Shape | Tooth_Enamel | Tecno | Tecno_type | Biped | Arm_morphology | Foot_motphology | Diet | Sexual_Dimorphism | Hip | Vertical_Front | Anatomy | Migrated | Skeleton |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Genus_Specie | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Location | 1 | 1 | 1 | 0.74 | 0.44 | 0.37 | 0.51 | 0.53 | 0.42 | 0.38 | 0.49 | 0.69 | 0.48 | 0.72 | 0.44 | 0.5 | 0.45 | 0.57 | 0.44 | 0.57 | 0.51 | 0.6 | 0.38 | 0.23 |
| Country | 1 | 1 | 1 | 0.61 | 0.6 | 0.47 | 0.48 | 0.58 | 0.52 | 0.59 | 0.57 | 0.6 | 0.61 | 0.59 | 0.52 | 0.61 | 0.55 | 0.61 | 0.55 | 0.62 | 0.63 | 0.65 | 0.63 | 0.59 |
| Habitat | 1 | 0.74 | 0.61 | 1 | 0.56 | 0.67 | 0.62 | 0.6 | 0.65 | 0.52 | 0.79 | 0.62 | 0.69 | 0.56 | 0.59 | 0.65 | 0.8 | 0.69 | 0.58 | 0.76 | 0.7 | 0.74 | 0.82 | 0.59 |
| Incisor_Size | 1 | 0.44 | 0.6 | 0.56 | 1 | 0.77 | 0.58 | 0.75 | 0.6 | 0.58 | 0.67 | 0.65 | 0.52 | 0.53 | 0.44 | 0.65 | 0.64 | 0.53 | 0.54 | 0.55 | 0.59 | 0.5 | 0.89 | 0.46 |
| Jaw_Shape | 1 | 0.37 | 0.47 | 0.67 | 0.77 | 1 | 0.7 | 0.88 | 0.74 | 0.62 | 0.64 | 0.8 | 0.53 | 0.53 | 0.61 | 0.55 | 0.84 | 0.73 | 0.49 | 0.63 | 0.6 | 0.62 | 0.48 | 0.55 |
| Torus_Supraorbital | 1 | 0.51 | 0.48 | 0.62 | 0.58 | 0.7 | 1 | 0.79 | 0.65 | 0.63 | 0.67 | 0.81 | 0.51 | 0.65 | 0.5 | 0.58 | 0.72 | 0.65 | 0.88 | 0.58 | 0.49 | 0.61 | 0.51 | 0.63 |
| Prognathism | 1 | 0.53 | 0.58 | 0.6 | 0.75 | 0.88 | 0.79 | 1 | 0.71 | 0.59 | 0.67 | 0.77 | 0.64 | 0.68 | 0.61 | 0.8 | 0.77 | 0.69 | 0.88 | 0.76 | 0.73 | 0.69 | 0.88 | 0.71 |
| Foramen_Mágnum_Position | 1 | 0.42 | 0.52 | 0.65 | 0.6 | 0.74 | 0.65 | 0.71 | 1 | 0.4 | 0.79 | 0.72 | 0.57 | 0.49 | 0.79 | 0.51 | 0.8 | 0.66 | 0.56 | 0.51 | 0.51 | 0.49 | 0.4 | 0.37 |
| Canine Size | 1 | 0.38 | 0.59 | 0.52 | 0.58 | 0.62 | 0.63 | 0.59 | 0.4 | 1 | 0.41 | 0.82 | 0.46 | 0.49 | 0.61 | 0.37 | 0.32 | 0.66 | 0.59 | 0.55 | 0.45 | 0.59 | 0.34 | 0.17 |
| Canines_Shape | 1 | 0.49 | 0.57 | 0.79 | 0.67 | 0.64 | 0.67 | 0.67 | 0.79 | 0.41 | 1 | 0.69 | 0.72 | 0.72 | 0.8 | 0.65 | 0.57 | 0.77 | 0.68 | 0.6 | 0.71 | 0.76 | 0.49 | 0.24 |
| Tooth_Enamel | 1 | 0.69 | 0.6 | 0.62 | 0.65 | 0.8 | 0.81 | 0.77 | 0.72 | 0.82 | 0.69 | 1 | 0.58 | 0.66 | 0.77 | 0.61 | 0.78 | 0.83 | 0.92 | 0.7 | 0.57 | 0.74 | 0.6 | 0.64 |
| Tecno | 1 | 0.48 | 0.61 | 0.69 | 0.52 | 0.53 | 0.51 | 0.64 | 0.57 | 0.46 | 0.72 | 0.58 | 1 | 1 | 0.66 | 0.65 | 0.71 | 0.67 | 0.6 | 0.64 | 0.6 | 0.61 | 0.68 | 0.37 |
| Tecno_type | 1 | 0.72 | 0.59 | 0.56 | 0.53 | 0.53 | 0.65 | 0.68 | 0.49 | 0.49 | 0.72 | 0.66 | 1 | 1 | 0.54 | 0.86 | 0.71 | 0.62 | 0.92 | 0.72 | 0.64 | 0.68 | 0.82 | 0.59 |
| Biped | 1 | 0.44 | 0.52 | 0.59 | 0.44 | 0.61 | 0.5 | 0.61 | 0.79 | 0.61 | 0.8 | 0.77 | 0.66 | 0.54 | 1 | 0.6 | 0.76 | 0.7 | 0.71 | 0.5 | 0.55 | 0.52 | 0.63 | 0.37 |
| Arm_morphology | 1 | 0.5 | 0.61 | 0.65 | 0.65 | 0.55 | 0.58 | 0.8 | 0.51 | 0.37 | 0.65 | 0.61 | 0.65 | 0.86 | 0.6 | 1 | 0.6 | 0.61 | 0.62 | 0.73 | 0.66 | 0.67 | 0.77 | 0.45 |
| Foot_motphology | 1 | 0.45 | 0.55 | 0.8 | 0.64 | 0.84 | 0.72 | 0.77 | 0.8 | 0.32 | 0.57 | 0.78 | 0.71 | 0.71 | 0.76 | 0.6 | 1 | 0.79 | 0.6 | 0.75 | 0.52 | 0.68 | 0.45 | 0.38 |
| Diet | 1 | 0.57 | 0.61 | 0.69 | 0.53 | 0.73 | 0.65 | 0.69 | 0.66 | 0.66 | 0.77 | 0.83 | 0.67 | 0.62 | 0.7 | 0.61 | 0.79 | 1 | 0.72 | 0.7 | 0.58 | 0.73 | 0.6 | 0.55 |
| Sexual_Dimorphism | 1 | 0.44 | 0.55 | 0.58 | 0.54 | 0.49 | 0.88 | 0.88 | 0.56 | 0.59 | 0.68 | 0.92 | 0.6 | 0.92 | 0.71 | 0.62 | 0.6 | 0.72 | 1 | 0.62 | 0.52 | 0.66 | 0.63 | 0.48 |
| Hip | 1 | 0.57 | 0.62 | 0.76 | 0.55 | 0.63 | 0.58 | 0.76 | 0.51 | 0.55 | 0.6 | 0.7 | 0.64 | 0.72 | 0.5 | 0.73 | 0.75 | 0.7 | 0.62 | 1 | 0.68 | 0.77 | 0.92 | 0.5 |
| Vertical_Front | 1 | 0.51 | 0.63 | 0.7 | 0.59 | 0.6 | 0.49 | 0.73 | 0.51 | 0.45 | 0.71 | 0.57 | 0.6 | 0.64 | 0.55 | 0.66 | 0.52 | 0.58 | 0.52 | 0.68 | 1 | 0.63 | 0.73 | 0.37 |
| Anatomy | 1 | 0.6 | 0.65 | 0.74 | 0.5 | 0.62 | 0.61 | 0.69 | 0.49 | 0.59 | 0.76 | 0.74 | 0.61 | 0.68 | 0.52 | 0.67 | 0.68 | 0.73 | 0.66 | 0.77 | 0.63 | 1 | 0.82 | 0.49 |
| Migrated | 1 | 0.38 | 0.63 | 0.82 | 0.89 | 0.48 | 0.51 | 0.88 | 0.4 | 0.34 | 0.49 | 0.6 | 0.68 | 0.82 | 0.63 | 0.77 | 0.45 | 0.6 | 0.63 | 0.92 | 0.73 | 0.82 | 1 | 0.34 |
| Skeleton | 1 | 0.23 | 0.59 | 0.59 | 0.46 | 0.55 | 0.63 | 0.71 | 0.37 | 0.17 | 0.24 | 0.64 | 0.37 | 0.59 | 0.37 | 0.45 | 0.38 | 0.55 | 0.48 | 0.5 | 0.37 | 0.49 | 0.34 | 1 |

- Certain categorical variables, such as Genus_Species, Habitat, and Diet, show strong correlations with other features:
- Genus_Species: Highly correlated with variables like Tecno, Tecno_type, and Habitat, indicating that these characteristics are closely linked to species classification.
- Diet: Strongly correlated with Canine Size and Tooth Enamel, reflecting specific dietary adaptations.
- Tecno_type: Moderately correlated with Biped, Diet, and Habitat, suggesting its relationship with anatomical and ecological changes.

# Correlation of Numerical Variables



- Very strong positive correlation: This suggests that species with larger cranial capacity also tend to have greater height.

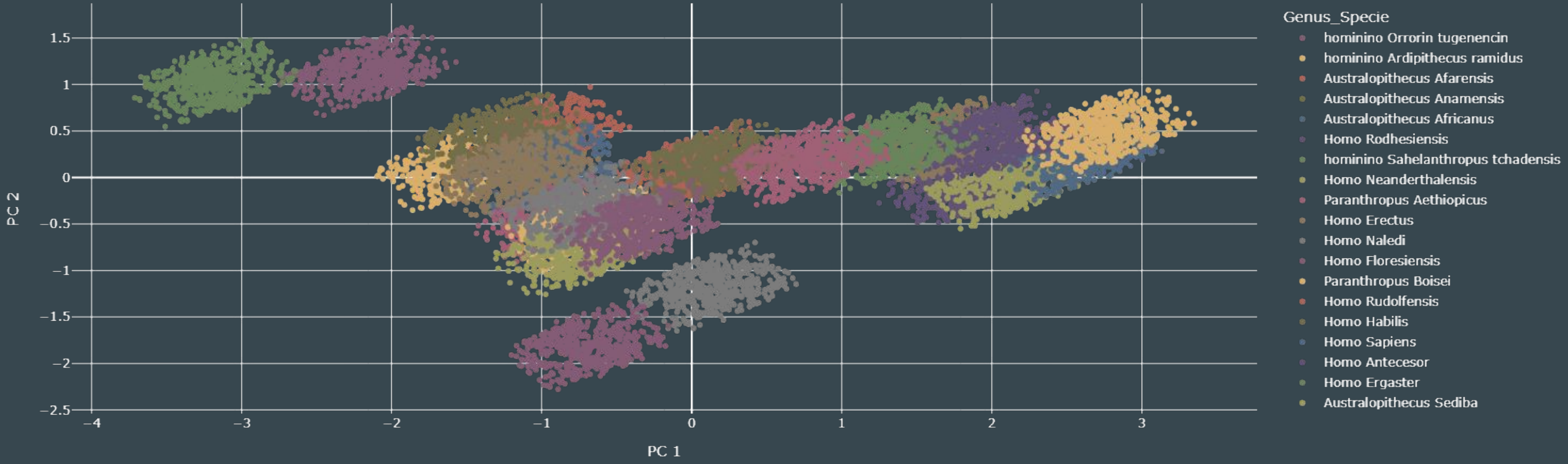
- Moderate negative correlation: More recent species tend to have a larger cranial capacity.

- Moderate negative correlation: More recent species also tend to be taller.

# PCA(Principal Component Analysis)

EThe graph presents a two-dimensional visualization of hominid species based on the first two Principal Components (PC1 and PC2), generated using Principal Component Analysis (PCA).

These components capture the majority of variability present in the selected numerical variables: Time, Cranial Capacity, and Height.



# Data Preprocessing

Scaling of Numerical Variables:

- Min-Max Scaler was used to scale numerical variables since the Shapiro-Wilks test indicated they did not follow a Gaussian distribution.
- This ensured that variable values were within a uniform range (0 to 1), optimizing their usability in the models.

Encoding of Categorical Variables:

- Ordinal Variables: Features with an inherent order, such as Canine Size and Incisor Size, were encoded using Ordinal Encoding, preserving their hierarchical structure.
- Non-Ordinal Variables: Categorical variables without a specific order, such as Habitat, Genus_Species, and Jaw Shape, were encoded using Label Encoding, assigning a unique number to each category without imposing hierarchies.

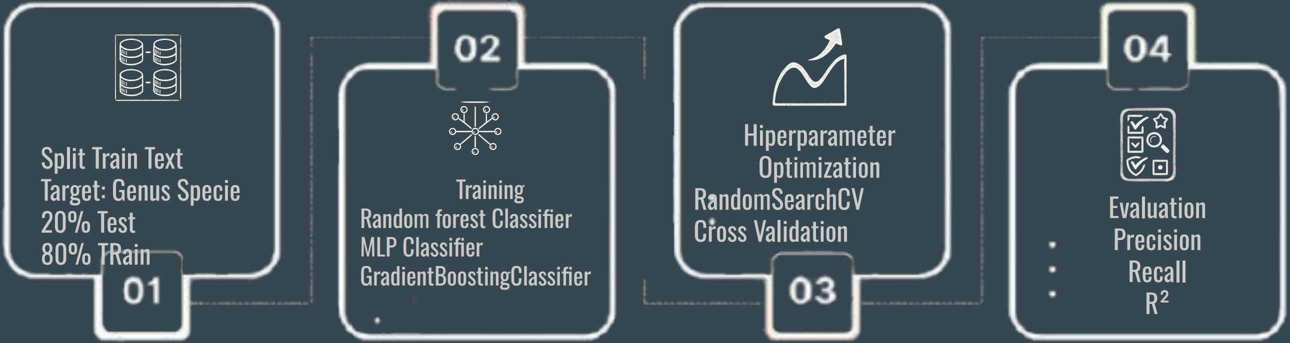
# Classification Models for Species Prediction

The Random Forest Classifier combines the results of multiple decision trees (through voting) to improve accuracy and reduce overfitting, leveraging the diversity of samples used in its construction.

It was chosen for its effectiveness in handling a mix of numerical and categorical variables, making it well-suited for this dataset.

The Multilayer Perceptron (MLP) is an artificial neural network designed to learn complex nonlinear patterns.

Its architecture includes multiple dense layers with ReLU activations, allowing it to capture deep relationships between the physical characteristics of the fossils.



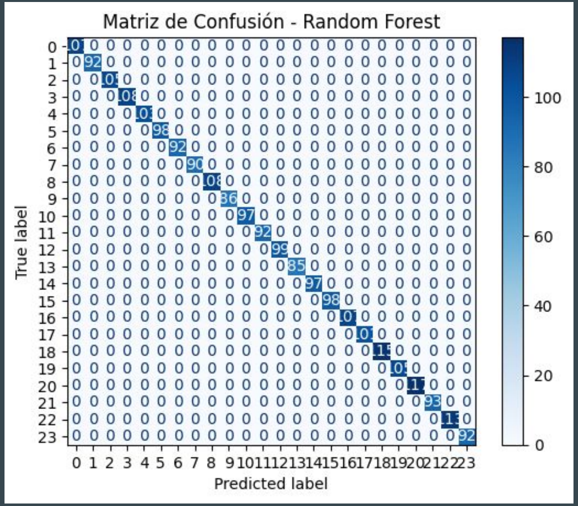
01 Split Train Text
Target: Genus Specie
20% Test
80% TRain

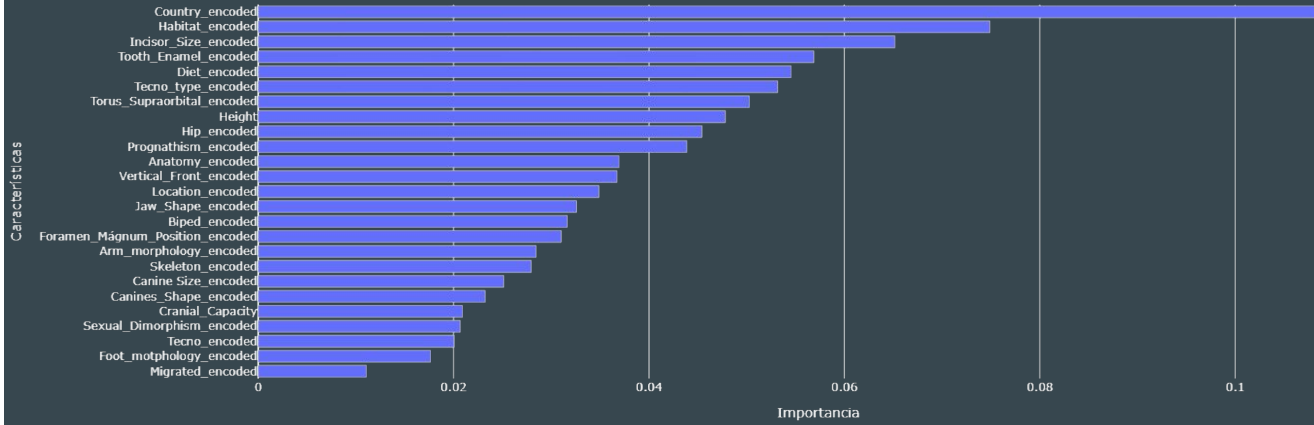
02 Training
Random forest Classifier
MLP Classifier
GradientBoostingClassifier

03 Hiperparameter Optimization
RandomSearchCV
Cross Validation

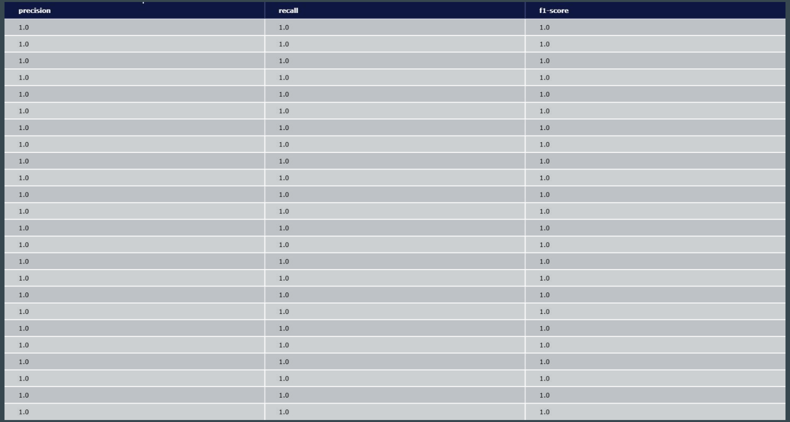
04 Evaluation
Precision
Recall
$R^2$

# Principales resultados de Random Forest



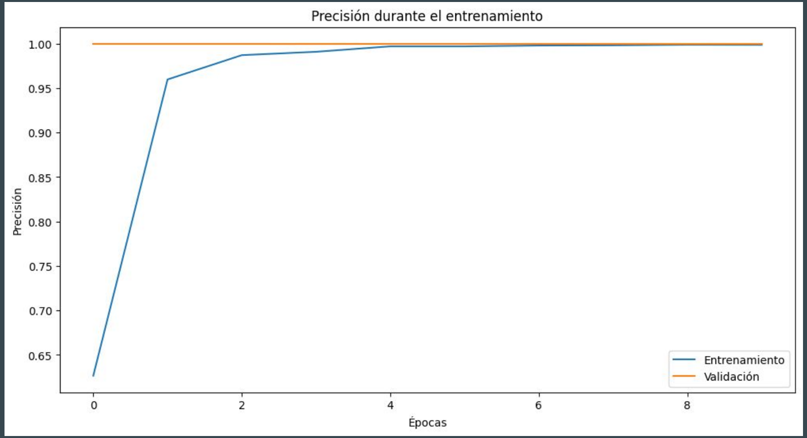


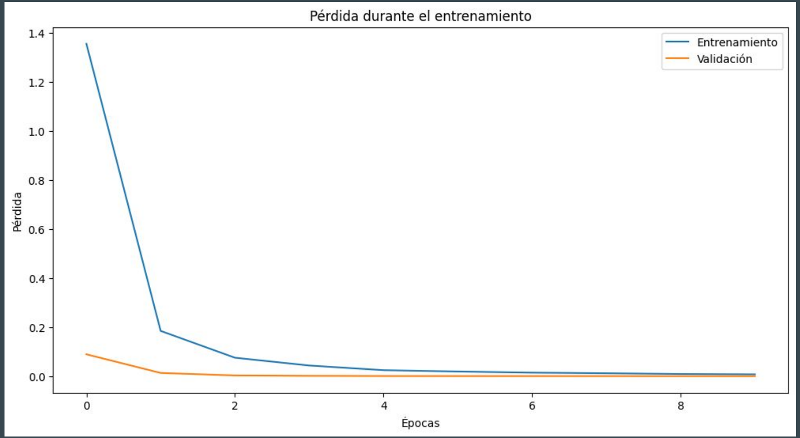
The model is classifying most instances correctly, as the majority of values are on the main diagonal.

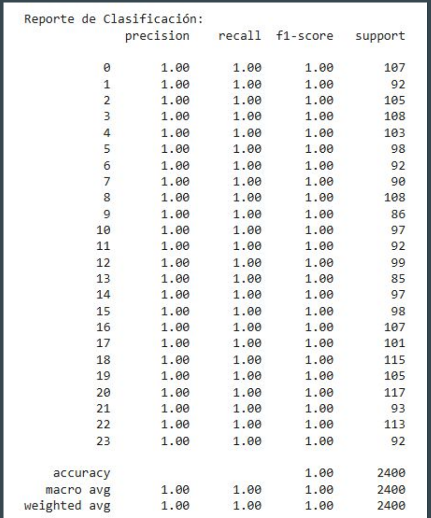


| precision | recall | f1-score |
|---|---|---|
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |
| 1.0 | 1.0 | 1.0 |

# Resultados de Multilayer Perceptron, MLP





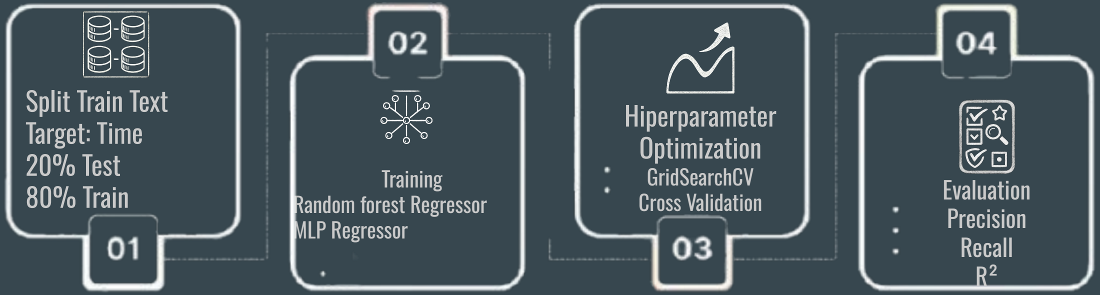


```
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       107
           1       1.00      1.00      1.00        92
           2       1.00      1.00      1.00       105
           3       1.00      1.00      1.00       108
           4       1.00      1.00      1.00       103
           5       1.00      1.00      1.00        98
           6       1.00      1.00      1.00        92
           7       1.00      1.00      1.00        90
           8       1.00      1.00      1.00       108
           9       1.00      1.00      1.00        86
          10       1.00      1.00      1.00        97
          11       1.00      1.00      1.00        92
          12       1.00      1.00      1.00        99
          13       1.00      1.00      1.00        85
          14       1.00      1.00      1.00        97
          15       1.00      1.00      1.00        98
          16       1.00      1.00      1.00       107
          17       1.00      1.00      1.00       101
          18       1.00      1.00      1.00       115
          19       1.00      1.00      1.00       105
          20       1.00      1.00      1.00       117
          21       1.00      1.00      1.00        93
          22       1.00      1.00      1.00       113
          23       1.00      1.00      1.00        92

    accuracy                           1.00      2400
   macro avg       1.00      1.00      1.00      2400
weighted avg       1.00      1.00      1.00      2400
```

- The model is achieving excellent performance in both training and validation.

- The low loss values in both sets suggest that the model is learning correctly and has a good fit.

# Modelos de regresión para datación preliminar



01

Split Train Text
Target: Time
20% Test
80% Train

02

Training

- Random forest Regressor
- MLP Regressor

03

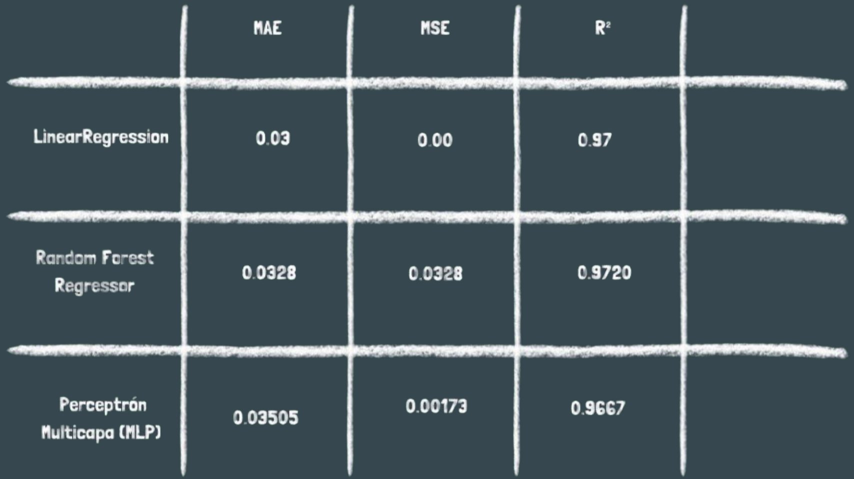
Hiperparameter Optimization

- GridSearchCV
- Cross Validation

04

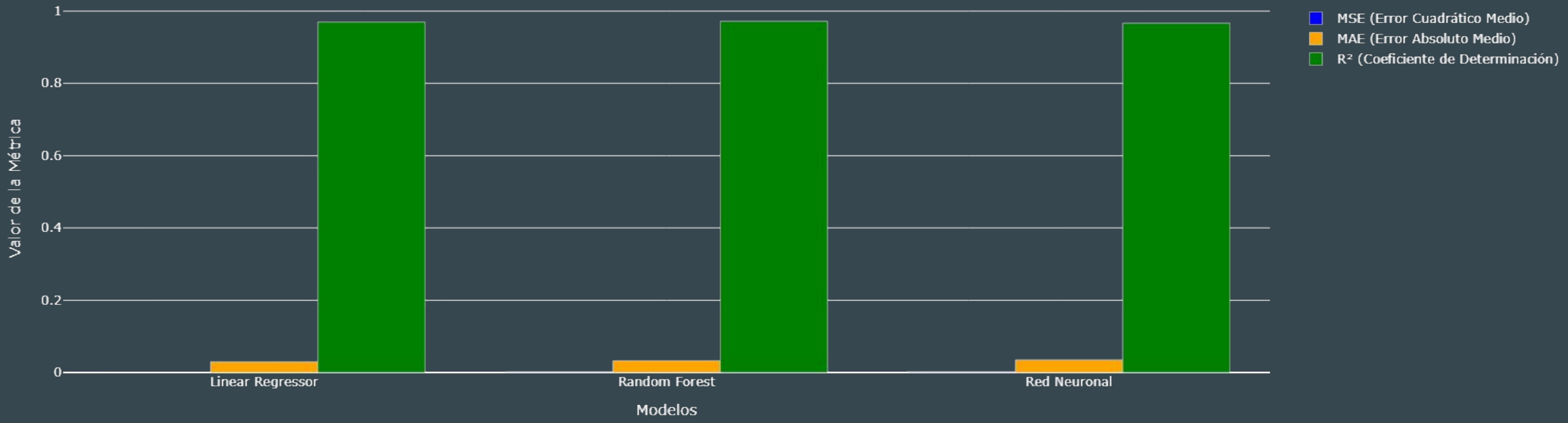
Evaluation

- Precision
- Recall
- $R^2$



| | MAE | MSE | $R^2$ |
|---|---|---|---|
| LinearRegression | 0.03 | 0.00 | 0.97 |
| Random Forest Regressor | 0.0328 | 0.0328 | 0.9720 |
| Perceptrón Multicapa (MLP) | 0.03505 | 0.00173 | 0.9667 |

# Model Comparison



# Conclusions

The development and implementation of the models demonstrated the feasibility of using data science for taxonomic classification.

The use of physical characteristics to predict genera and species suggests that these variables contain consistent patterns that can be leveraged for classification purposes. This enables faster and more accurate taxonomic identification.

The models have proven effective in estimating geological time with minimal errors, which can be useful in contexts where absolute dating is unavailable.

The choice of model depends on the objective:

Taxonomic classification: Random Forest.

Time estimation: Linear regression for simple cases, neural networks or Random Forest for more complex relationships.

# Limitations and Future Work

Limitations

Subjectivity in Measurement: Some anatomical characteristics can be difficult to quantify objectively, introducing a certain degree of subjectivity in the data.

Taxonomic Criteria: The definition of species and genus in paleontology is a complex and evolving subject. Classification criteria may vary among researchers, making it challenging to compare results.

Future Work

Expanding the Dataset: Incorporating new fossil samples, particularly those with uncommon traits or from understudied geographic regions.

Establishing a Reproducible Framework: Creating a methodology that can be replicated for future data-driven studies. This approach could also be transferred to forensic anthropology for predicting characteristics in modern individuals.