UNIVERSITATEA BABES BOLYAI, CLUJ NAPOCA, ROMANIA
FACULTATEA DE MATEMATICA SI INFORMATICA

Reconstituiri istorice

– MIRPR –

Membrii echipei

Andrei-Danut Blagoi, Informatica romana, grupa 231

Andreea Bolonyi, Informatica romana, grupa 231

Stefan-Nicolae Parvanescu, Informatica engleza, grupa 936

Oana-Alexandra Sidorencu, Informatica romana, grupa 236

2021-2022

Rezumat

Proiectul propus este menit sa vina in ajutorul persoanelor pasionate de istorie si arheologie care isi doresc informatii plastice despre anumite date introduse. Utilizatorii aplicatiei au posibilitatea sa exploreze niste date numerice pe care le gasesc, reusind sa cunoasca care este sexul sau varsta unui os pe care acestia il studiaza si, de asemenea, utilizatorii au posibilitatea sa perceapa informatia si intr-un mod vizual.

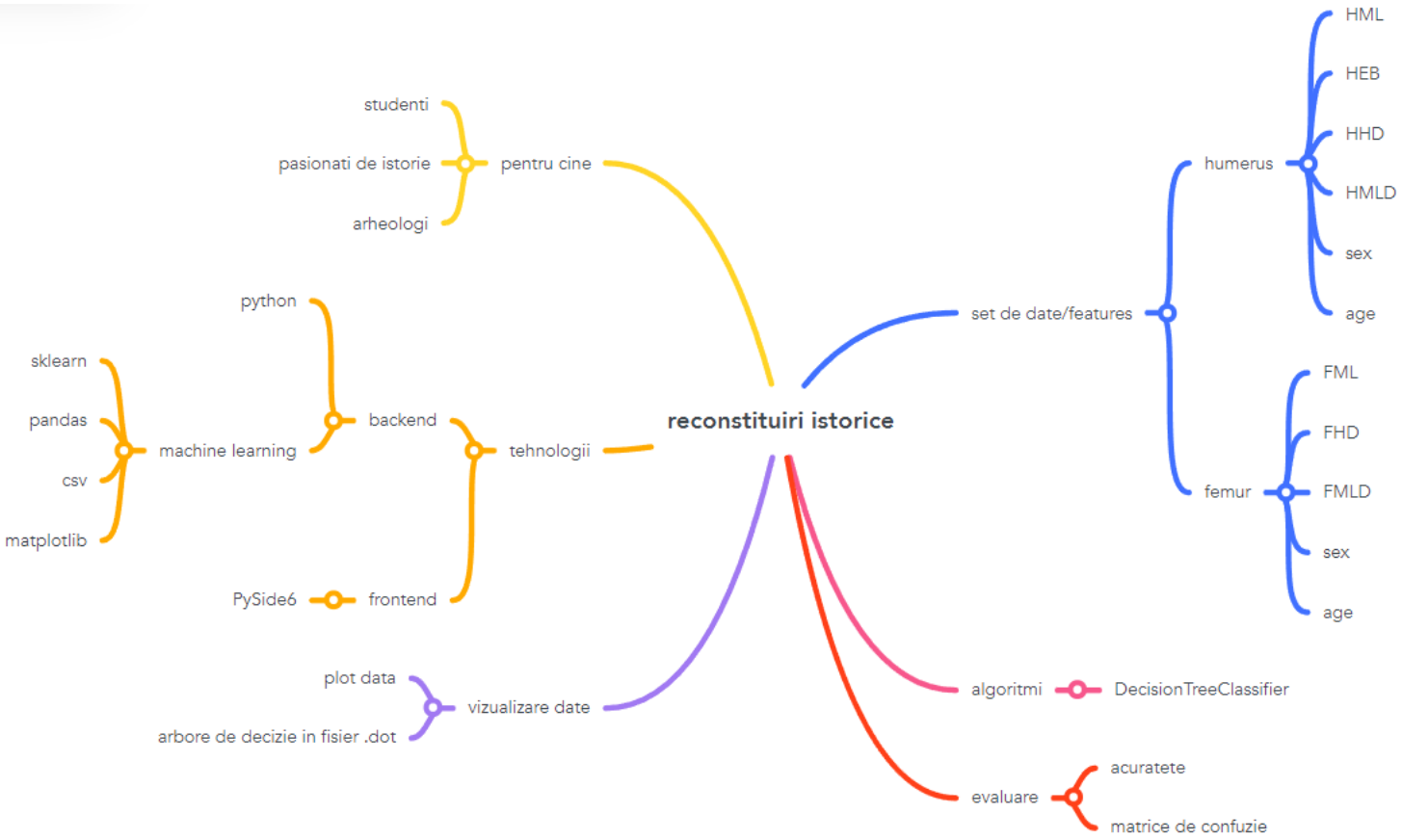

Figura 1: Mind map.

Cuprins

Capitolul 1

Introducere

1.1 Motivarea temei

Problema abordata in acest proiect este de interes pentru persoanele care lucreaza in acest domeniu, studenti ce vor ajunge angajati sau persoane care sunt doar pasionate. Aplicatia dezvoltata este usor de folosit, intuitiva si menita sa ofere doar ajutor, nu probleme utilizatorului.

Pasionatii folosind aplicatia noastra vor putea sa aiba o idee mult mai aprofundata in legatura cu subiectul pe care il studiaza, nu doar sa se bazeze pe niste cifre pe care le vad. De asemenea, o functionalitate pe care dorim sa o oferim utilizatorului este posibilitatea de a vedea 3D aceste informatii pe care le introduce.

1.2 Tehnologii

Dezvoltarea aplicartiei se bazeaza pe limbajul Python pentru partea de backend si PyQT pentru partea de frontend. Am decis sa folosim Python datorita usurintei cu care se poate scrie codul si multitudinea de solutii pe care le putem gasi, fie ca este vorba de o problema obisnuita pentru un programator, fie ca este vorba de biblioteci din domeniul inteligentei artificiale pe care le putem folosi pentru rezolvarea subiectului (spre exemplu Scikit-learn, TensorFlow, Theano).

1.3 Setul de date

S-a folosit un set de date de aproximativ 2000 de exemple pentru humerus din care aproximativ 1000 sunt barbati si aproximativ 1000 sunt femei.

TML	Tibia Maximum Length
TPB	Tibia Plateau Mediolateral (Bicondylar) Breadth
TMLD	Tibia 50% Diaphyseal Mediolateral Diameter
TAPD	Tibia 50% Diaphyseal Anteroposterior Diameter

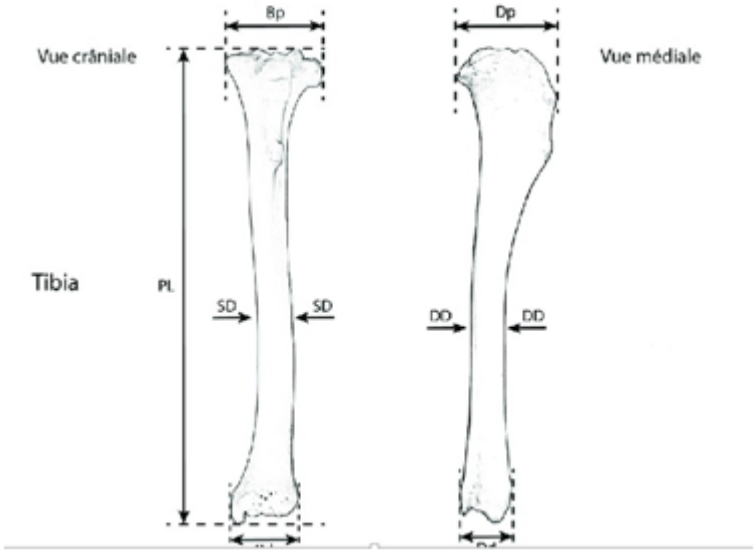

Figura 1.1: Caracteristici tibie.

HML	Humerus Maximum Length –b (33,4)M (30,7)F
HEB	Humerus Epicondylar Breadth -c
HHD	Humerus Head Diameter -g
HMLD	Humerus 50% Diaphyseal Mediolateral Diameter -a
HAPD	Humerus 50% Diaphyseal Anteroposterior Diameter -a

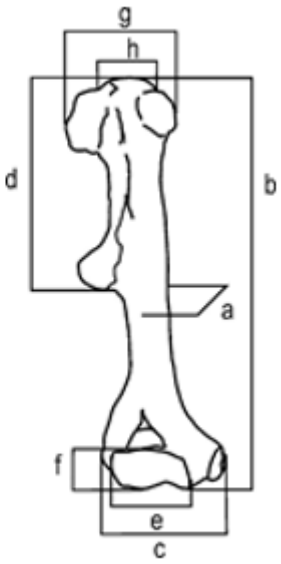

Figura 1.2: Caracteristici humerus.

RML	Radius Maximum Length
RMLD	Radius 50% Diaphyseal Mediolateral Diameter (MAX) -a
RAPD	Radius 50% Diaphyseal Anteroposterior Diameter (MIN) -a

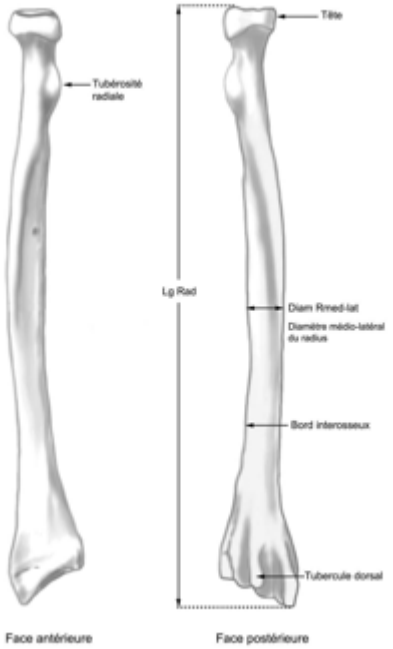

Figura 1.3: Caracteristici radius.

FML	Femur Maximum Length –FML
FBL	Femur Bicondylar Length - FBL
FEB	Femur Epicondylar Mediolateral Breadth – FEB
FAB	
FHD	
FMLD	Femur 50% Diaphyseal Mediolateral Diameter
FAPD	Femur 50% Diaphyseal Anteroposterior Diameter

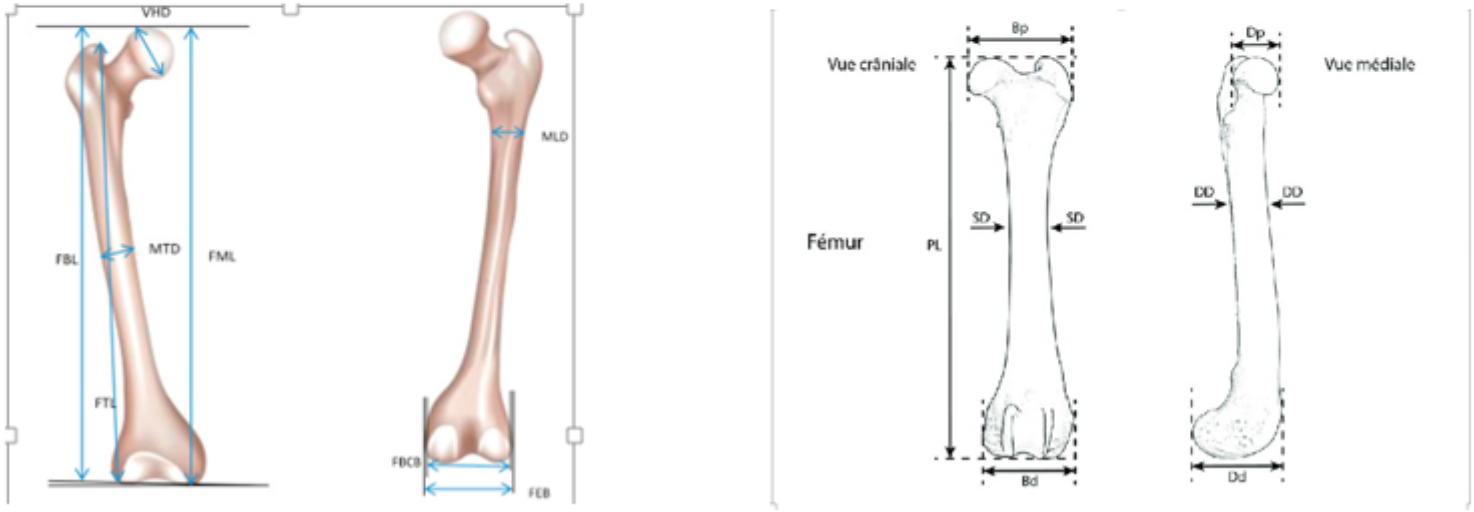

Figura 1.4: Caracteristici femur.

Capitolul 2

Problema stiintifica

2.1 Definitia problemei

Subiectul abordat tine de domeniul istoriei si al arheologiei pentru a obtine informatii revelante despre obiectele identificate in santierele arheologice. Astfel se doreste o aplicatie care, plecand de la informatii deja studiate de arheologi umani, sa permita vizualizarea 3D a unor "descoperiri deja efectuate" in intregime sau partial, din diferite unghiuri, reliefand anumite detalii. Mai mult, ofera posibilitatea determinarii sexului sau varstei pe baza unor caracteristici numerice ale unui anumit tip de os.

Pentru simplitatea aplicatiei si usurinta folosirii exista o interfata grafica care va permite utilizatorului sa introduca caracteristicile pe baza carora se va stabili rezultatul. Dupa apasarea unui buton de trimitere a datelor, utilizatorul va fi intrebat daca este de acord ca datele introduse sa fie pastrate in baza de date pentru imbunatatirea solutiei. In urma procesarii datelor, utilizatorul va vedea care este sexul sau varsta osului specificat si va avea posibilitatea sa observe si in 3D cum ar arata acesta.

2.2 Algoritm machine learning - arbore de decizie

Ca prim algoritm am folosit un arbore de decizie datorita simplitatii de a intelege cum mai exact se iau deciziile pentru a determina clasele din care face parte o instanta; este usor de inteles pentru o persoana din domeniul informaticii, dar si pentru orice alta persoana.

Pentru un arbore de decizie avem un nod radacina, noduri intermediare si noduri finale(frunze). Nodul radacina reprezinta atributul cel mai semnificativ, nodurile intermediare care reprezinta decizii de

tip daca-atunci si frunzele care reprezinta clasa din care face parte o instanta. Pentru a determina rezultatul final practic se imparte setul de date in instante care respecta un anumit set de reguli.

Selectarea unui atribut ca fiind nod radacina se poate face folosind mai multi indici. Indexul Gini este varianta care se foloseste de clasa din Sklearn (DecisionTreeClassifier) daca nu este specificat alt index. Acesta este o metrica care masoara cat de des un element ales aleator este clasificat gresit; un index Gini mai mic inseamna un atribut care va fi preferat pentru a deveni nod radacina. In implementarea aplicatiei s-a folosit de asemenea un index Gini.

2.3 Rezultate arbore de decizie

O prima varianta de implementare pentru problema curenta de clasificare a folosit clasa din Sklearn (DecisionTreeClassifier) cu indexul Gini pentru a stabili nodul radacina. S-a folosit un set de date de dimensiune 100 care avea distirbutia descrisa ulterior. De asemenea, s-a creat un arbore in care se poate observa cum un nod intermediar reprezinta o decizie care se poate lua, ea fiind de tipul "daca conditia este adevarata, atunci mergi pe partea stanga a subarborelui, altfel pe partea dreapta; daca avem o frunza atunci ne oprim si returnam clasa corespunzatoare".

Folosind aceasta implementare am observat rezultate bune, avand o acuratete pentru determinarea sexului (probabilitatea ca un exemplu sa fie clasificat corect) de aproximativ 80% cu un timp de executie redus pentru antrenarea arborelui. Totusi, un dezavantaj ce a fost remarcat a fost faptul ca arborele este destul de sensibil la modificarea datelor de intrare si o mica eroare de tipar poate da peste cap algoritmul(de exemplu, daca arborele primeste o instante care are numele caracteristicilor cu alt format fata de cum este dat in structura lui s-ar putea sa aiba probleme la asocierea valorilor cu sensul lor).
Initial acuratetea era de aproximativ 65% atunci cand aveam setul de date initial cu peste 2000 de exemple, dar dupa ce am echilibrat setul de date si am ajuns aproximativ la celasi numar de femei si de barbati acuratetea a crescut la 80%.

Am folosit pentru determinarea varstei tot un arbore de decizie, avand drept clase urmatoarele:

- mai putin de 20 de ani
- intre 20 si 30 de ani
- intre 30 si 40 de ani

- intre 40 si 50 de ani
- peste 50 de ani

In acest caz am observat o acuratete destul de mica initial, in jur de 35-40% si un arbore mult mai stufos si greu de parcurs. Timpul de executie nu se poate spune ca s-a marit sau nu, rezultatele se obtin destul de repede.Exact ca in cazul determinarii sexului, setul de date are aproximativ 1900 de exemple.

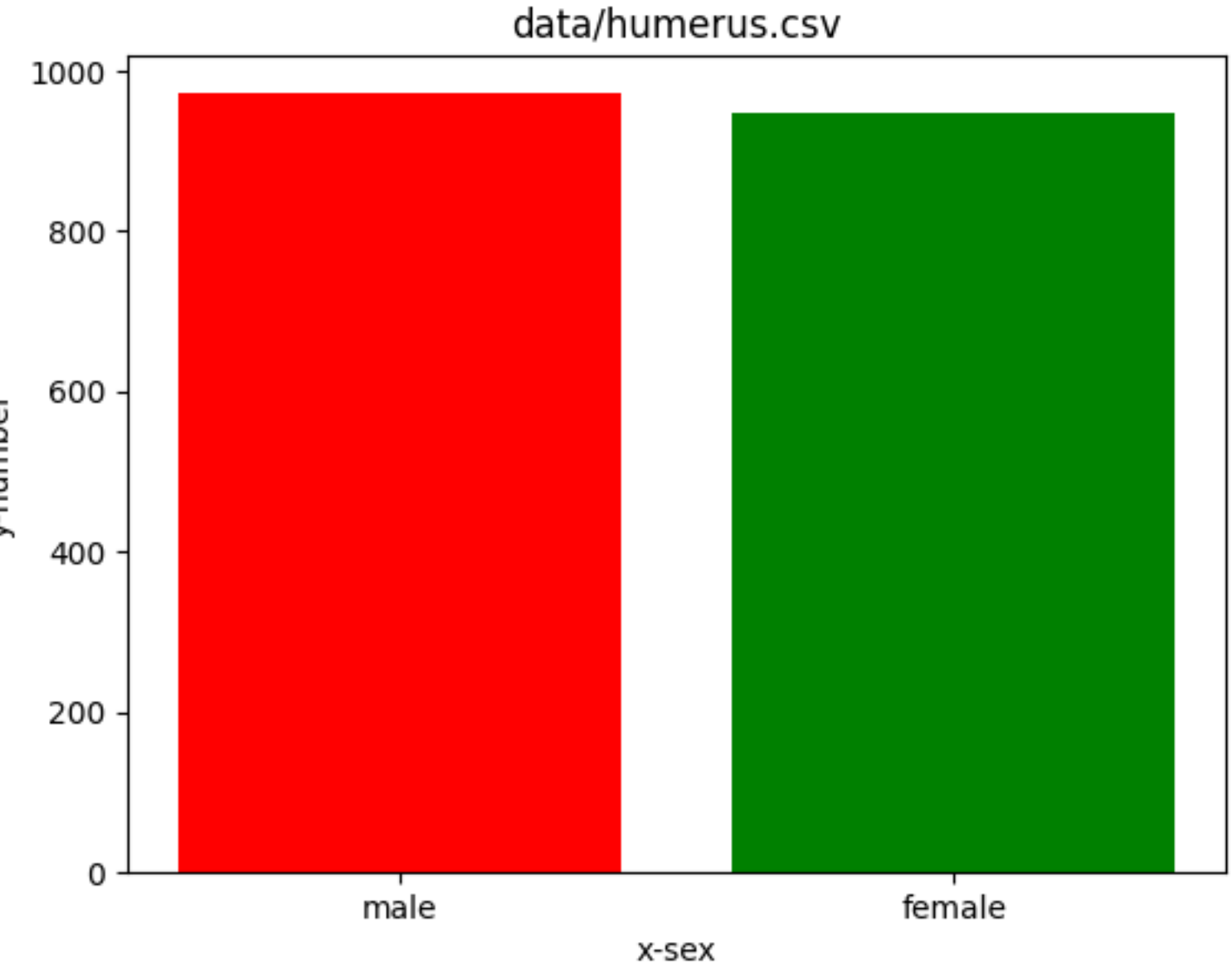

Figura 2.1: Distributie date in functie de sex.

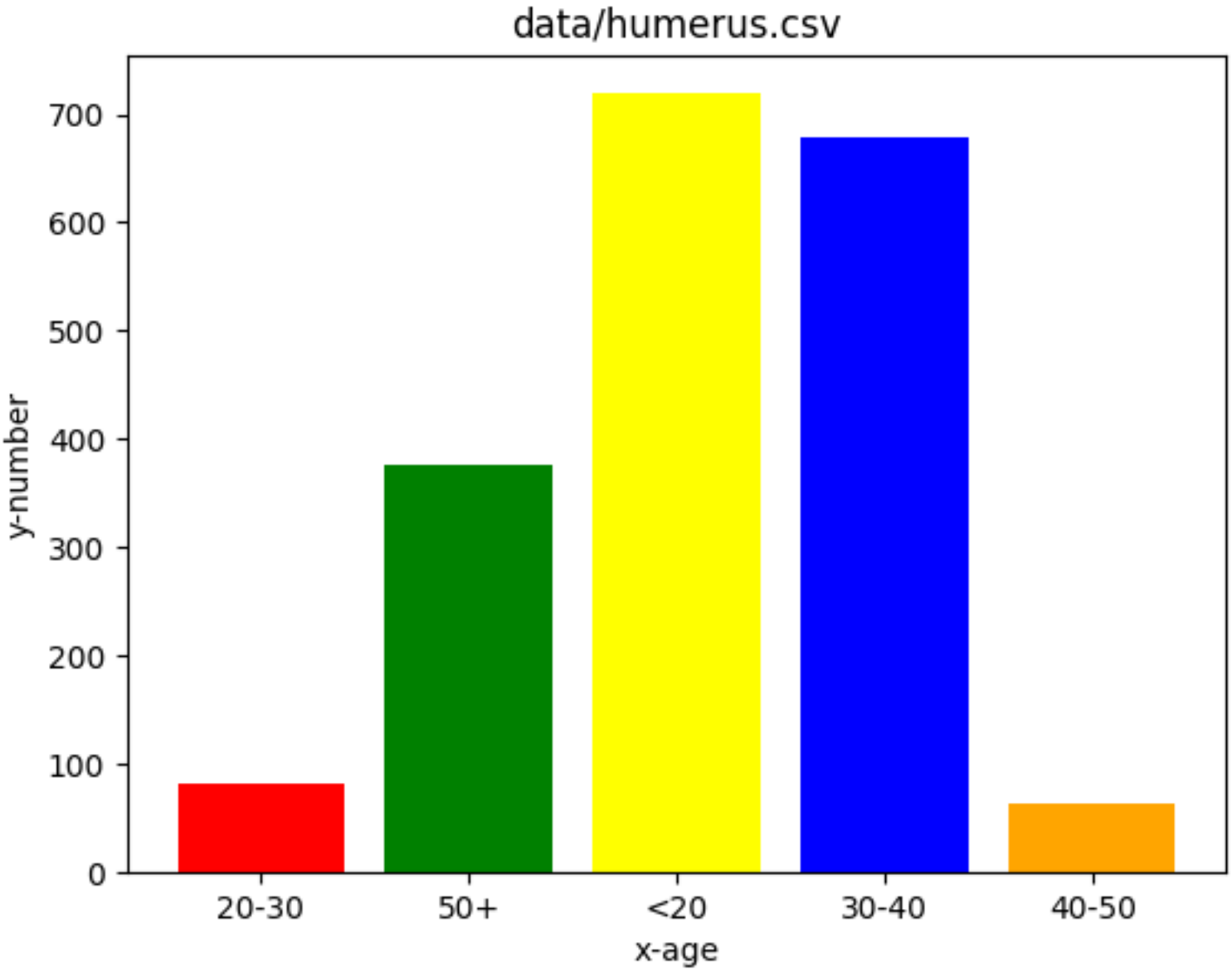

Figura 2.2: Distributie date in functie de varsta.

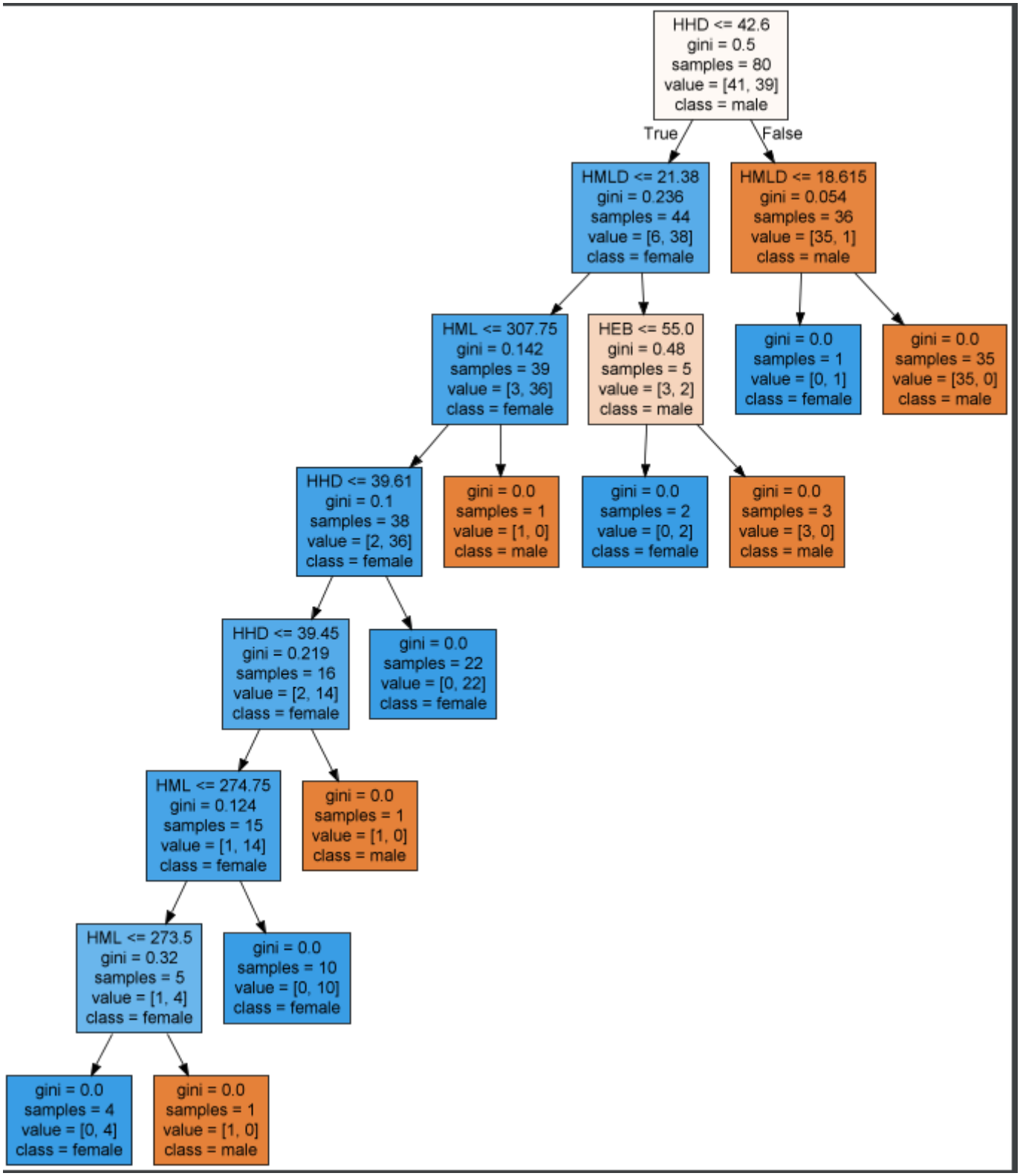

Figura 2.3: Arbore generat pentru un set de date cu 100 de exemple, clasificare sex.

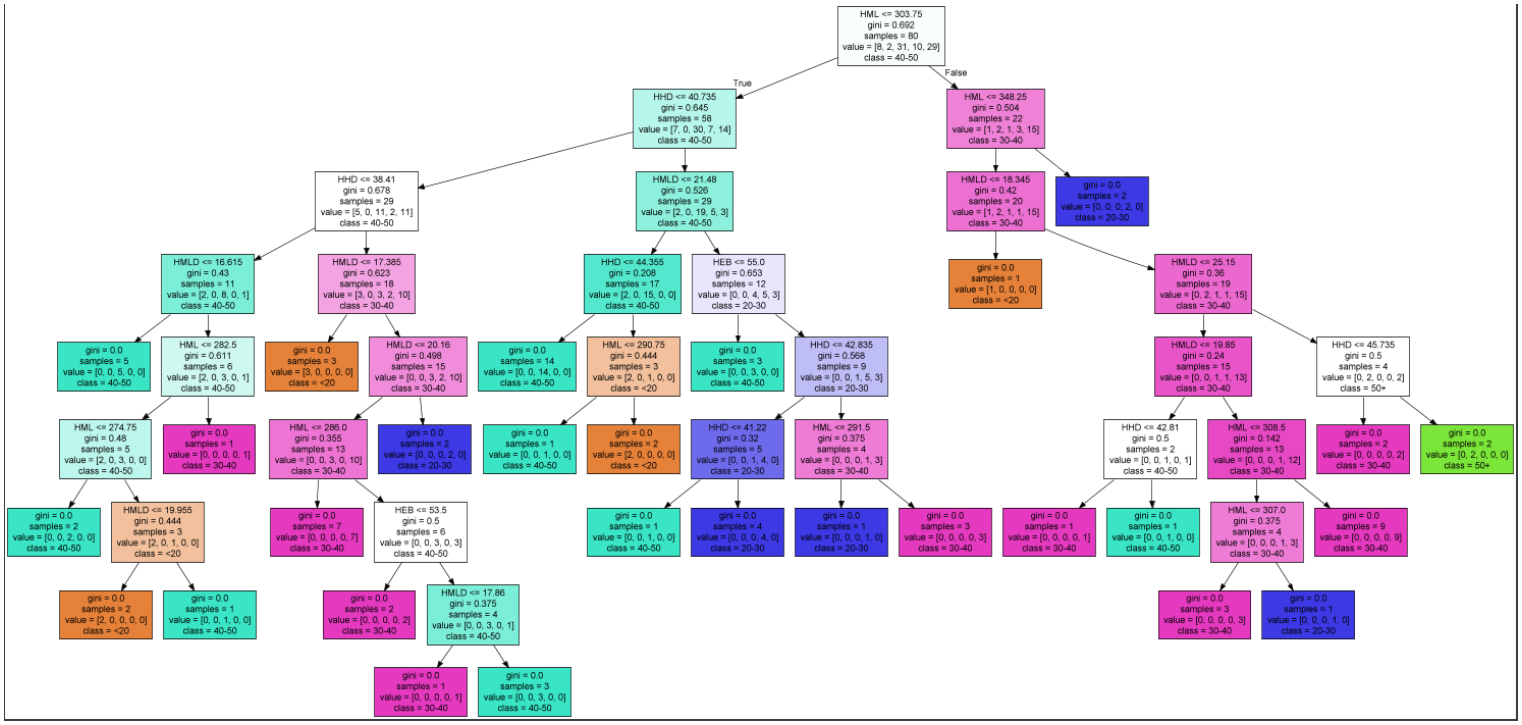

Figura 2.4: Arbore generat pentru un set de date cu 100 de exemple, clasificare varsta.

Capitolul 3

SOTA (State Of The Art)

3.1 Arbore de decizie

Lucrarea stiintifica care a stat la baza alegerii de a folosi arbori de decizie pentru a determina varsta a fost *SEX IDENTIFICATION IN ARCHAEOLOGICAL REMAINS USING DECISION TREE LEARNING*, avandu-i ca autori pe Ioan-Gabriel Mircea, Gabriela Czibula si Mara-Renata Petrusel, an 2015.

S-a folosit algoritimul ID3 (Iterative Dichotomiser 3) care foloseste un set de date S pe care il considera nod radacina. La fiecare iteratie a algoritmului se itereaza prin fiecare atribut nefolosit din S si se calculeaza entropia acelui atribut. Se selecteaza atributul cu cea mai mica entropie si astfel se imparte setul de date in mai multe subseturi. Algortimul se poate opri in unul din cazurile:

- fiecare element din subset apartine aceleasi clase, caz in care nodul este transformat in frunza si etichetat cu clasa exemplelor din subset
- nu mai exista atribute ce pot fi selectate si exemplele nu fac parte din aceeasi clasa; in acest caz nodul este transformat in frunza si este etichetat cu clasa cea mai comuna din exemplele subsetului
- nu mai exista exemple in subset care se intampla in momentul in care niciun exemplu din setul initial nu a fost gasit sa i se potriveasca o valoare cu atributul selectat; in acest caz se creeaza un nod frunza care este etichetat cu clasa cea mai comuna a exemplelor din setul initial

Setul de date cu care s-a lucrat a continut 200 de barbati si 200 de femei. In primul caz de testare un os a fost caracterizat de 10 masuratori legate de radius, in al doilea caz au fost 9 caracteristici ce tineau de antebrat, iar al treilea caz a fost reuniunea datelor din primele doua cazuri, deci s-a ajuns la 19 caracteristici pentru antrebrat si radius.

Acuratetea cea mai buna obtinuta a fost de 86% pentru primul caz, pentru al doilea a fost 87% si pentru al treilea s-a reusit o acuratete de 88%.

Astfel, comparand cu rezultatele obtinute cu aborele de decizie folosit in aceasta aplicatie se poate observa ca diferenta nu este atat de mare si acuratetea se apropie de ceea ce s-a obtinut in articolul stiintific. Una din diferente este faptul ca s-au folosit 200 de barbati si 200 de femei in lucrare, iar in aplicatia curenta s-au folosit aproximativ 1000 de barbati si aproximativ 1000 de femei. A doua diferenta consta in numarul de caracteristici ce s-au folosit pentru antrenarea algoritmului. In cazul aplicatiei curente s-au folosit 4 masuratori ale unui humerus (HML - humerus maximum length,HEB - humerus epicondylar breadth, HHD - humer head diameter, HMLD - humerus 50% diaphyseal mediolateral diameter).

Capitolul 4

Wiki

Sklearn

- Arbore de decizie
- Acuratete
- Matrice de confuzie
- Vizualizare arbore de decizie
- Export arbore de decizie in fisier .dot (decizii ca text)

Citire CSV(Comma Separated Values)

- Pandas
- csv tool

Set de date

- Goldman Osteometric Data Set
- Goldman guide to the measurements
- European Data Set-May 2018

Vizualizare date

- Matplotlib
- Bar chart

Machine learning

- Algoritm arbore de decizie articol stiintific mentionat drept referinta
- Implementare arbori de decizie din sklearn
- Implementare arbori de decizie din sklearn a doua varianta

Capitolul 5

Lucrari stiintifice

Punctul de start al proiectului dat a fost constituit de urmatoarele lucrari stiintifice: [2], [5], [6], [4], [1], [7], [3]. Acestea au oferit inspiratia si ajutorul de care am avut nevoie pentru a dezvolta aplicatia.

Pentru implementarea arborelui de decizie si alegerea acestuia in aplicatie s-a folosit lucrarea [3], de unde am retinut mai ales importanta alegerii unui algoritm care sa fie usor de inteles atat pentru programator, cat si pentru un arheolog pentru a verifica corectitudinea deciziilor. De asemenea, pentru alegerea metricii care selecteaza atributul s-au luat in considerare informatiile prezentate in [7].

Bibliografie

[1] Gabriela Czibula. *Machine learning-based approaches for predicting stature from archaeological skeletal remains using long bone lengths.* 2016.

[2] Geertje Klein Goldewijk and Jan Jacobs. *The relation between stature and long bone length in the roman empire.* 2013.

[3] Gabriela Czibula si Mara-Renata Petrusel Ioan-Gabriel Mircea. *SEX IDENTIFICATION IN ARCHAEOLOGICAL REMAINS USING DECISION TREE LEARNING.* 2015.

[4] Gabriela Czibula Ionescu, Vlad-Sebastian and Mihai Teletin. *Supervised Learning Techniques for Body Mass Estimation in Bioarchaeology.* 2016.

[5] Jan PAM Jacobs Jongman, Willem M. and Geertje M. Klein Goldewijk. *Health and wealth in the Roman Empire.* 2019.

[6] Diana-Lucia Miholca. *Machine learning based approaches for sex identification in bioarchaeology.* 2016.

[7] Dr. K. Nirmala R. Aruna devi. *Construction of Decision Tree : Attribute Selection Measures.* 2013.