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.

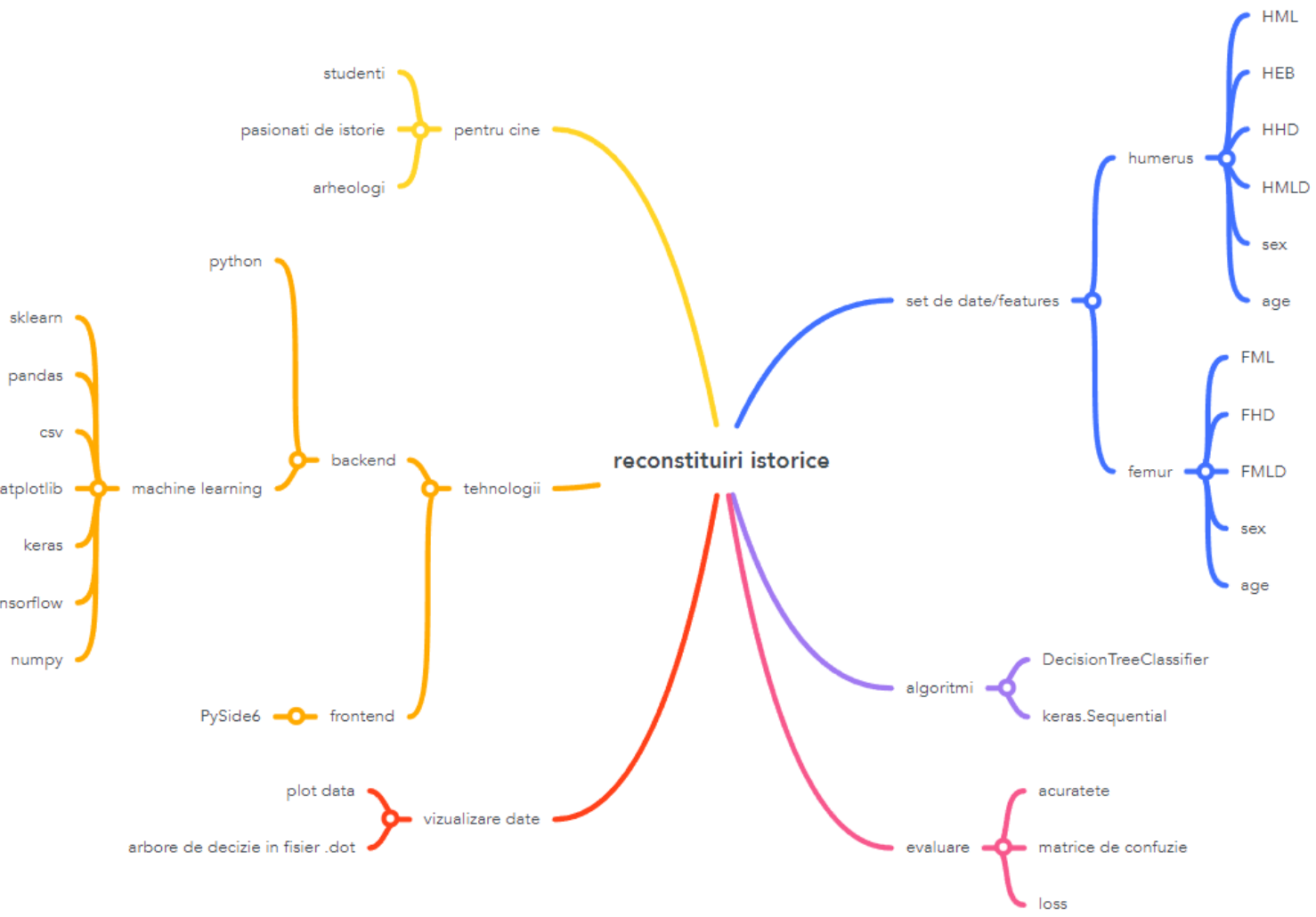

Figura 1: Mind map.

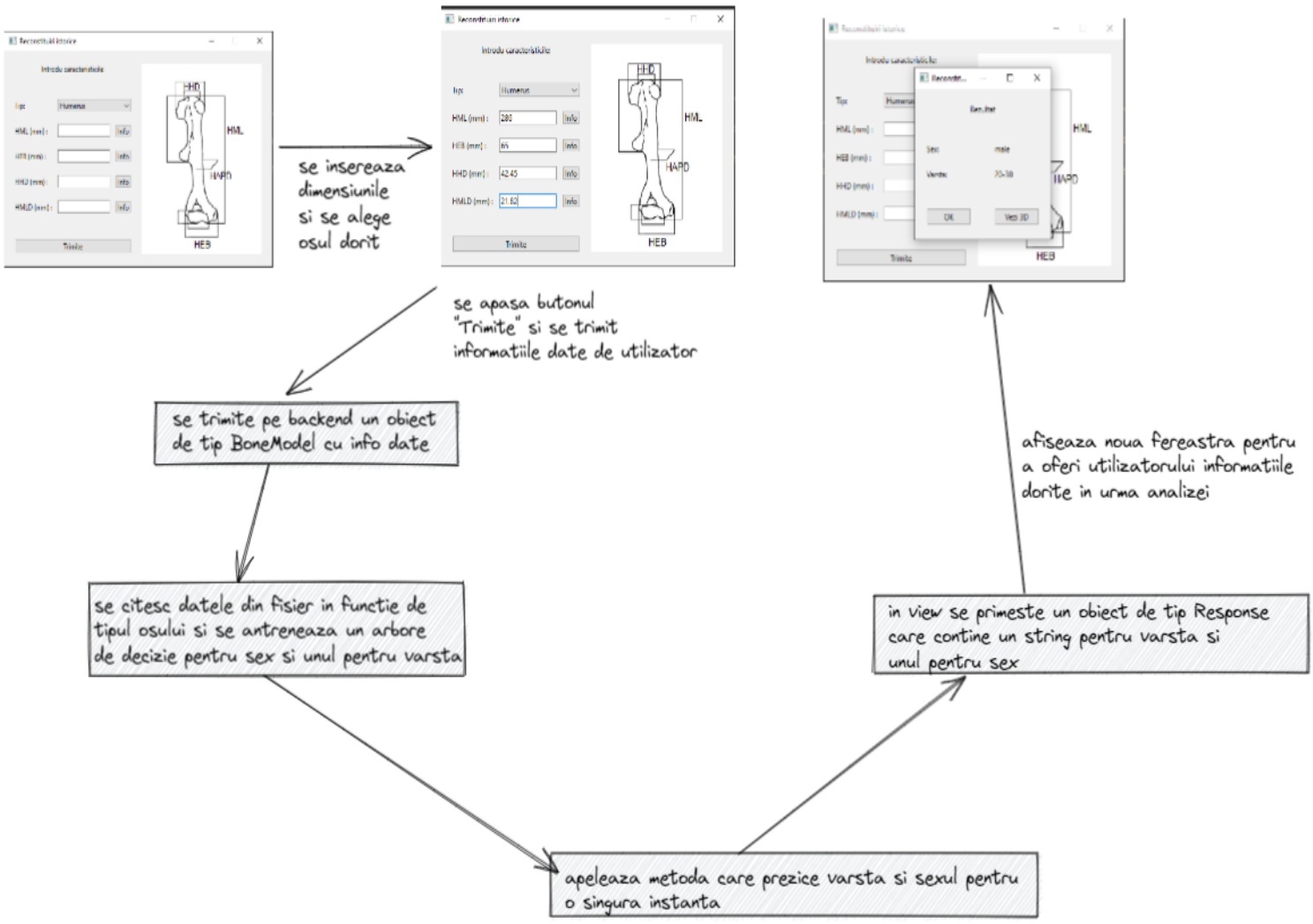

Figura 2: Flow al aplicatiei.

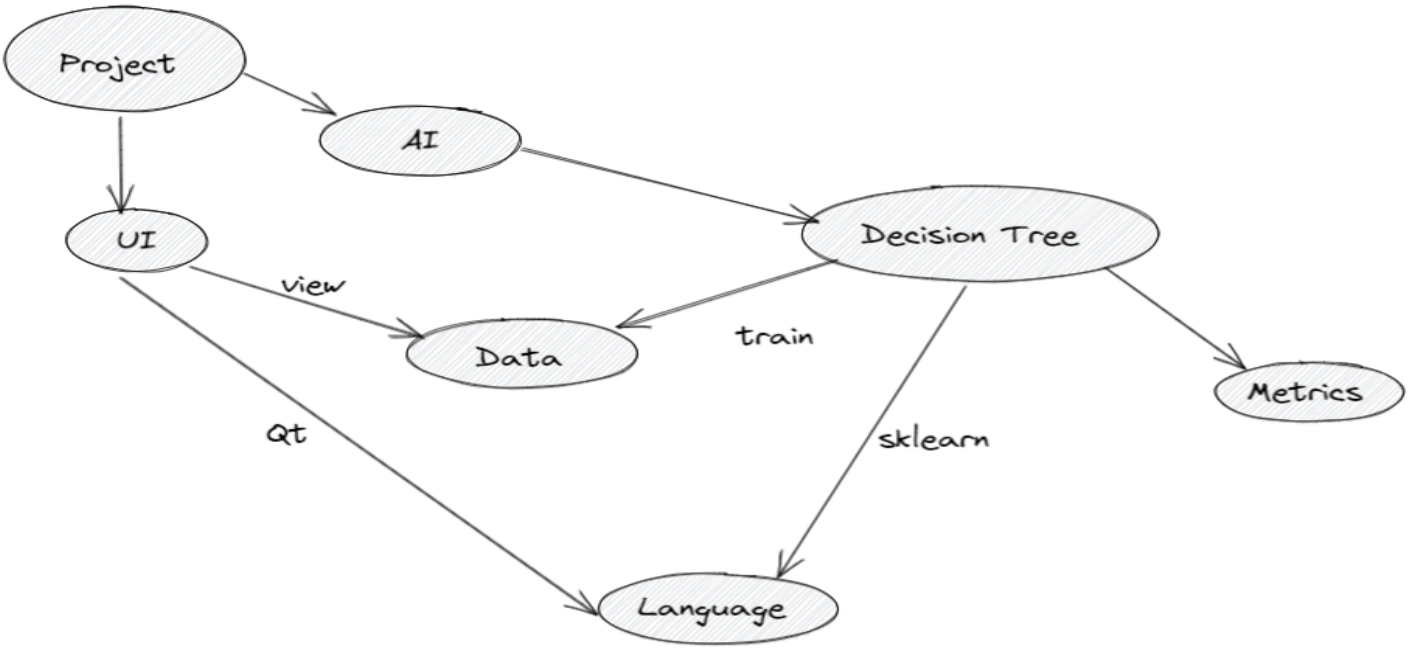

Figura 3: Ontologia pentru arborele de decizie.

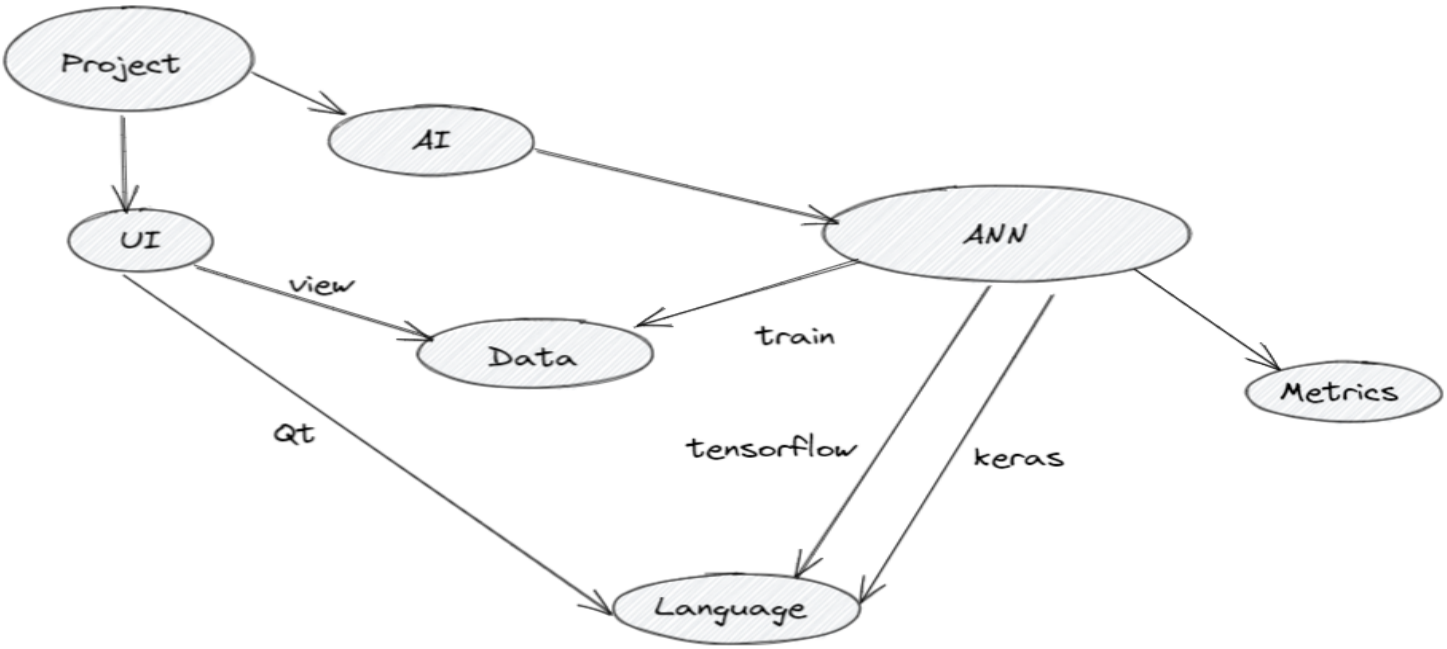

Figura 4: Ontologia pentru reteaua neuronala.

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 Masuratori ale oaselor

Caracteristicile folosite pentru fiecare tip de os si semnificatiile lor.

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

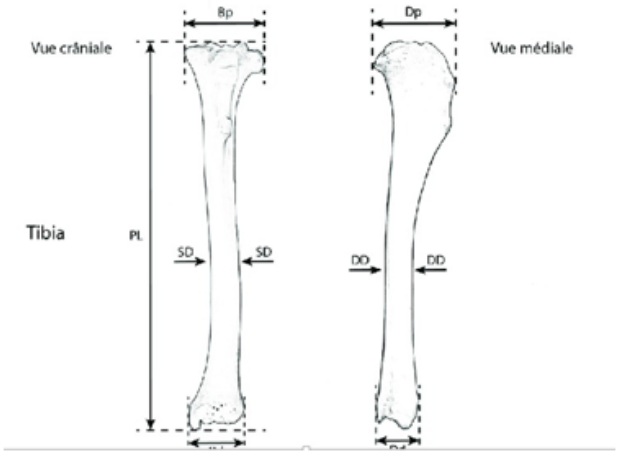

Figura 1.1: Caracteristici tibie.

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

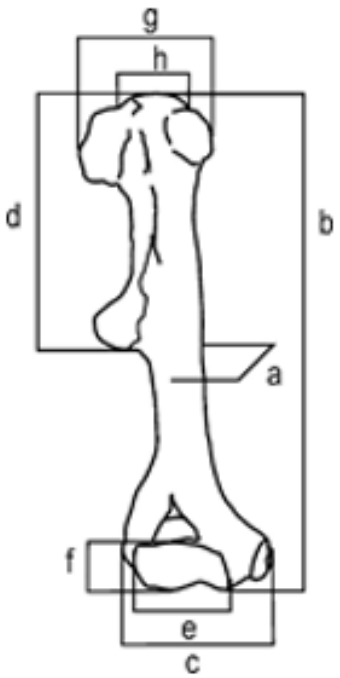

Figura 1.2: Caracteristici humerus.

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

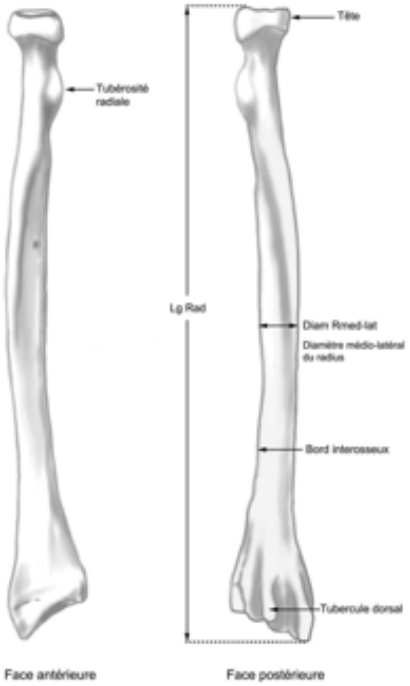

Figura 1.3: Caracteristici radius.

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

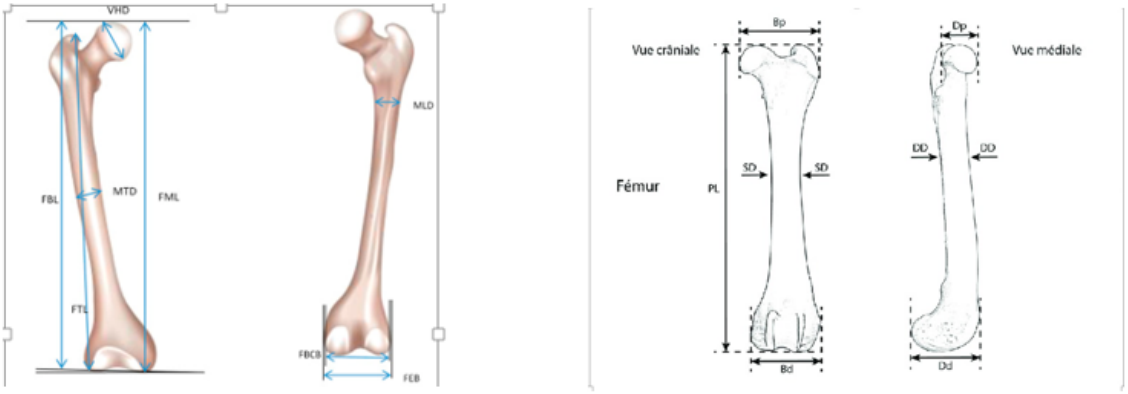

Figura 1.4: Caracteristici femur.

1.4 Setul de date

Setul de date cu care s-a lucrat pentru fiecare tip de os (reprezentare grafica a numarului de elemente din fiecare categorie).

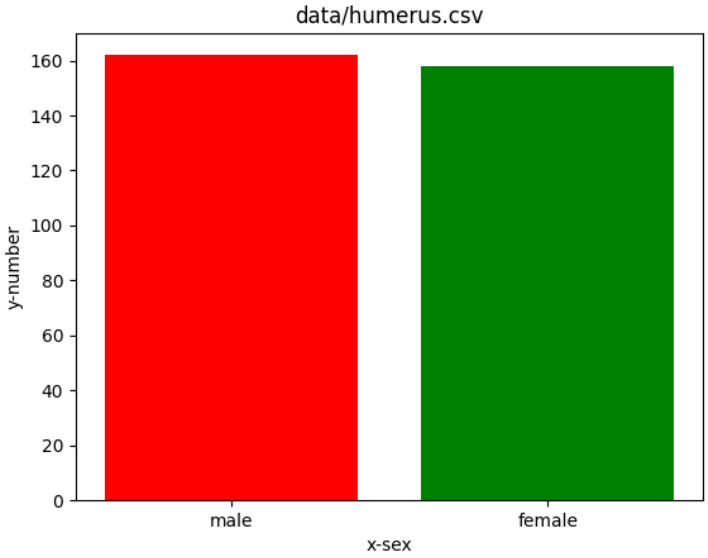

Figura 1.5: Distributie dupa sex pentru humerus.

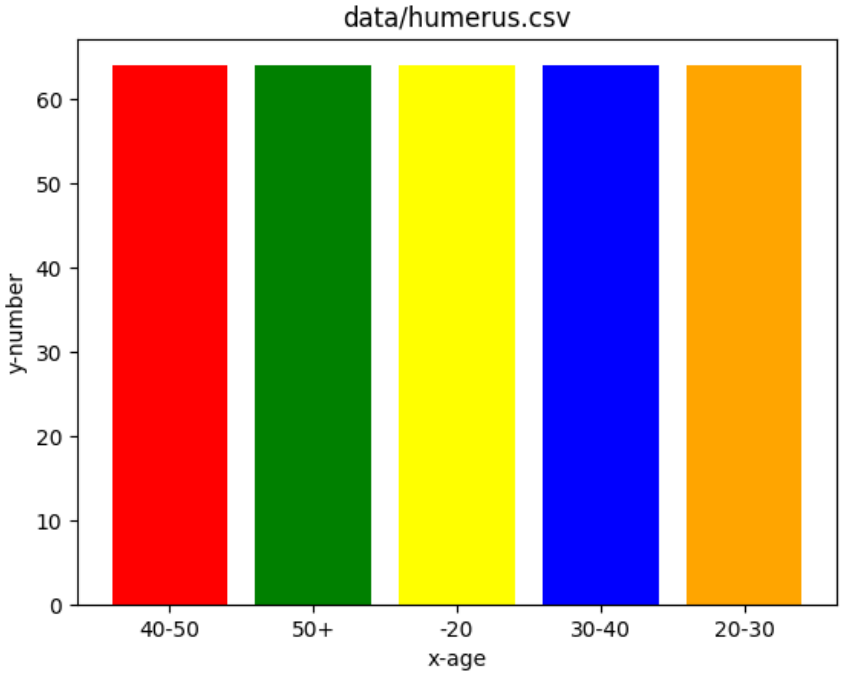

Figura 1.6: Distributie dupa varsta pentru humerus.

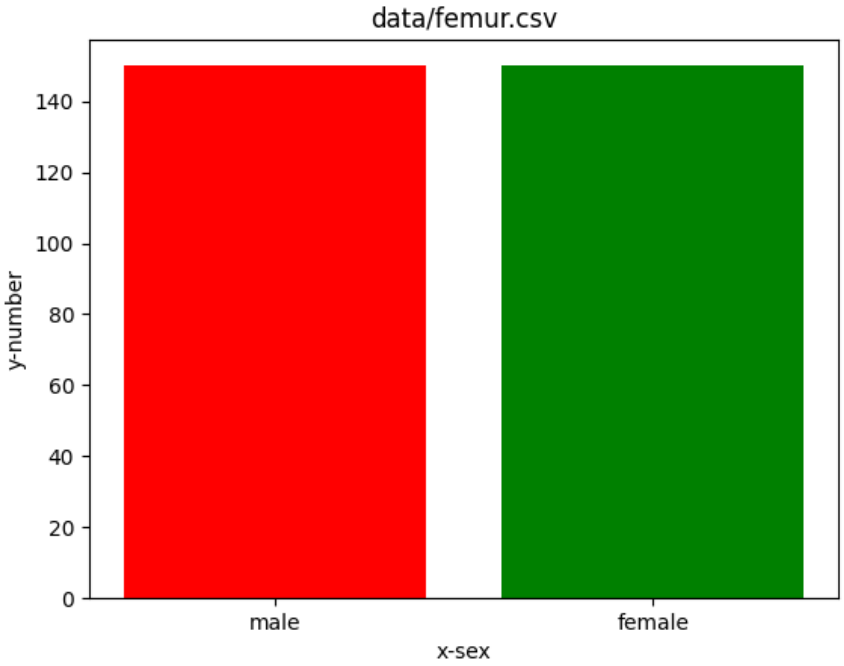

Figura 1.7: Distributie dupa sex pentru femur.

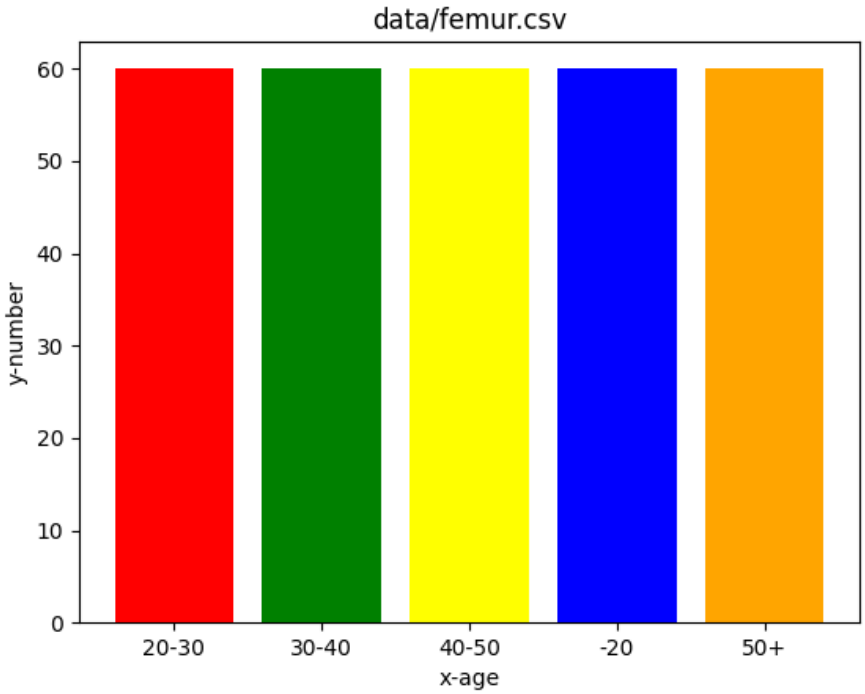

Figura 1.8: Distributie dupa varsta pentru femur.

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 30% 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. Dupa echilibrarea setului de date s-a ajuns la o acuratete de aproximativ 45%, ajungand la un numar mult mai redus de exemple cu care este antrenat algoritmul (aproximativ 300).

2.4 Algoritm machine learning - retele neuronale

Ca un al doilea algoritm am folosit o retea neuronala. O retea neuronala este alcatuita din o multime de noduri (neuroni) dispuse ca un graf pe mai multe straturi (layere). Nodurile au rolul de a efectua calculi simple prin intermediul unei functii associate (functie de activare) si sunt conectate intre ei prin niste legaturi ponderate. Straturile sunt de 3 tipuri: Strat de intrare care contine m neuroni unde m reprezinta numarul de caracteristici al unui obiect, Stratul de iesire care contine r neuroni,unde r reprezinta numarul de subclase si strat intermediar.

2.5 Rezultate retele neuronale

Pentru rezolvarea problemei de clasificare a sexului am impartit datele ca fiind 80% din ele date de antrenament si 20% pentru date de testare.
Reteaua neuronala este alcaturita din:

- stratul de intrare ce contine 4 neoroni
- strat intermediar de 20 de neoroni
- strat intermediar de 40 de neuroni
- stratul de iesire cu 2 neoroni

Pentru Humerus acuratetea modelului de clasificare de sex se afla intre 80-87% in timp ce pentru femur acuratetea este intre 80-83%.

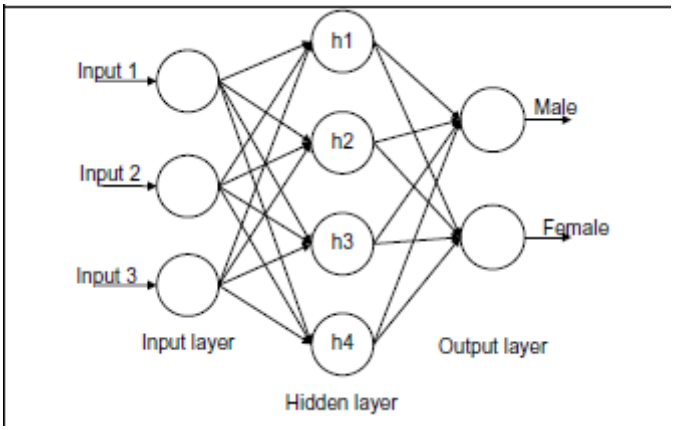

Figura 2.1: Structura retelei neuronale pentru clasificarea in functie de sex.

Humerus	Training	Validation	Femur	Training	Validation
1	81.48	87.24	1	81.55	83.98
2	78.47	87.76	2	80.79	82.14
3	79.01	87.76	3	83.19	82.69
Media	79.65	87.58	Media	81.84	82.93

Figura 2.2: Rezultate obtinute pentru mai multe testari.

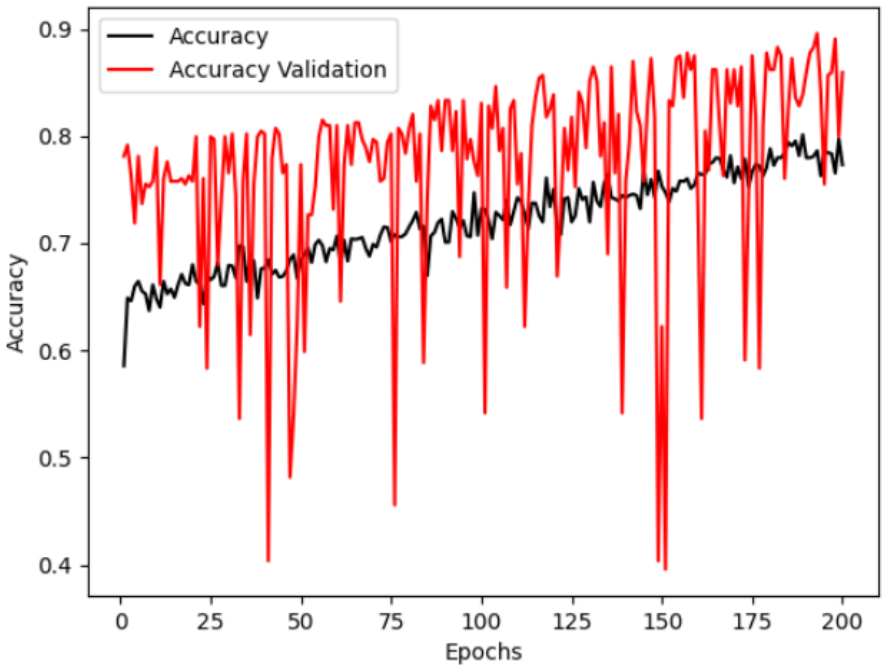

Figura 2.3: Acuratetea obtinuta pentru humerus in functie de numarul de epoci.

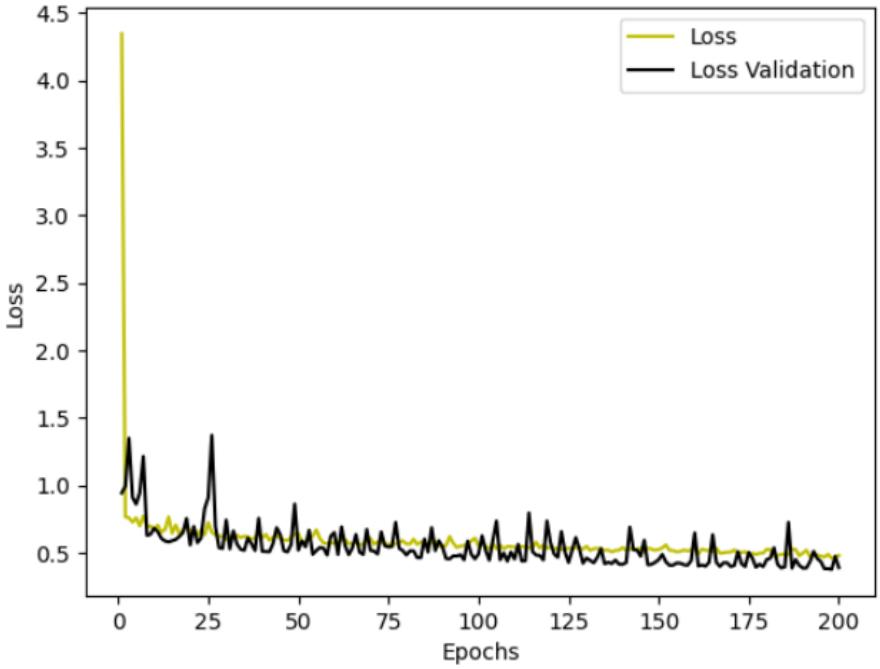

Figura 2.4: Loss obtinut pentru humerus in functie de numarul de epoci.

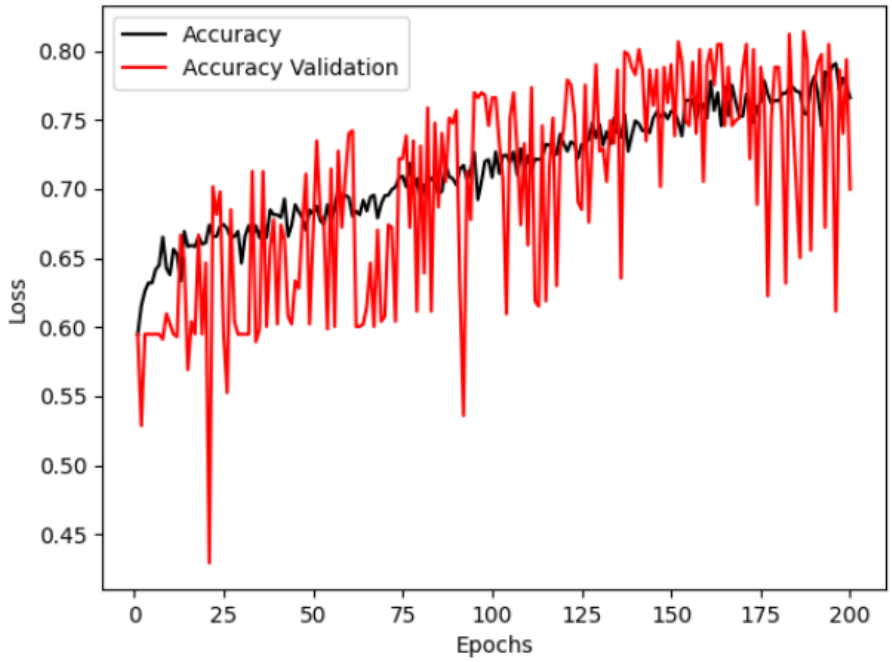

Figura 2.5: Acuratetea obtinuta pentru femur in functie de numarul de epoci.

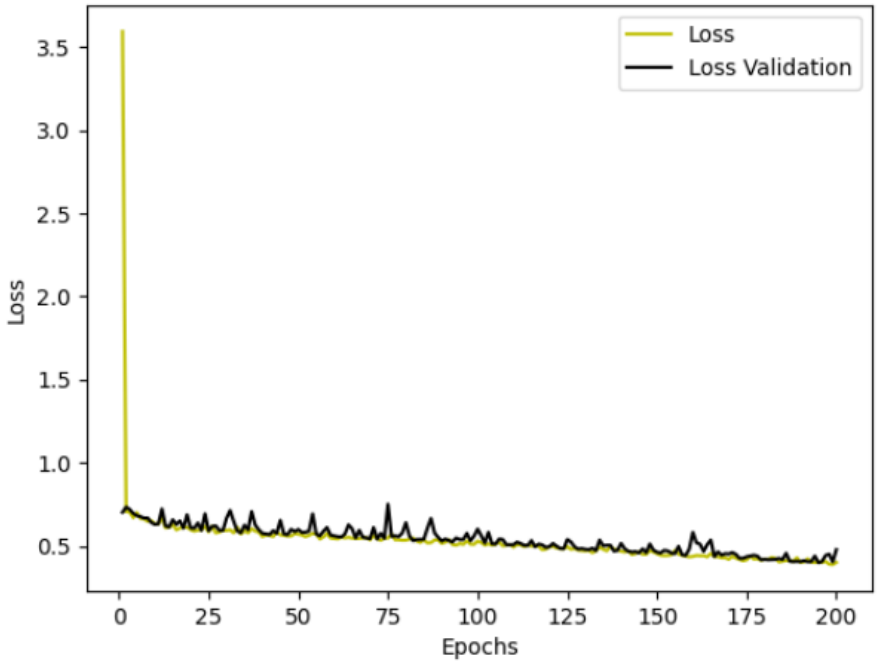

Figura 2.6: Loss obtinut pentru femur in functie de numarul de epoci.

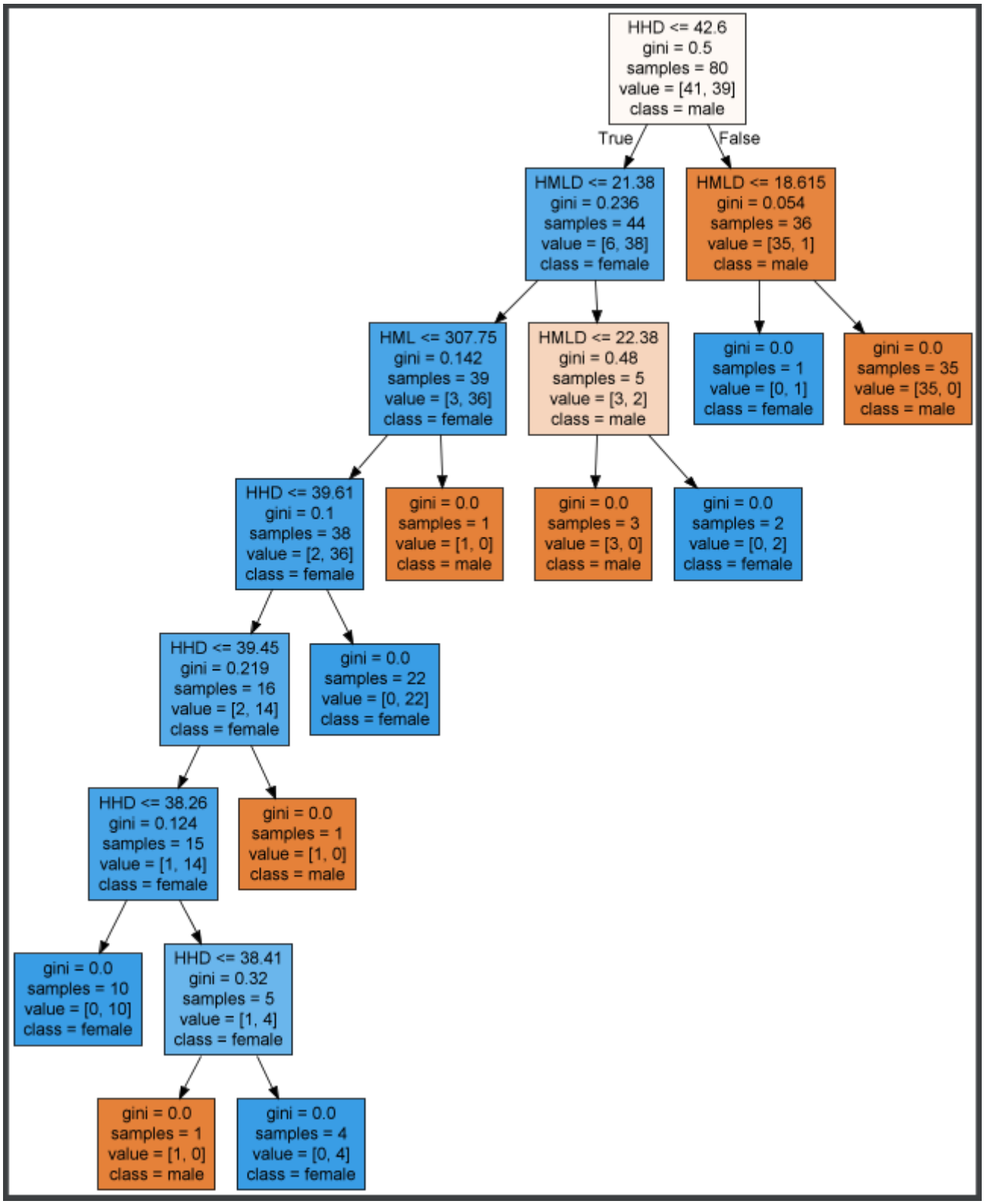

Figura 2.7: Arbore generat pentru un set de date cu 100 de exemple, clasificare sex pentru humerus.

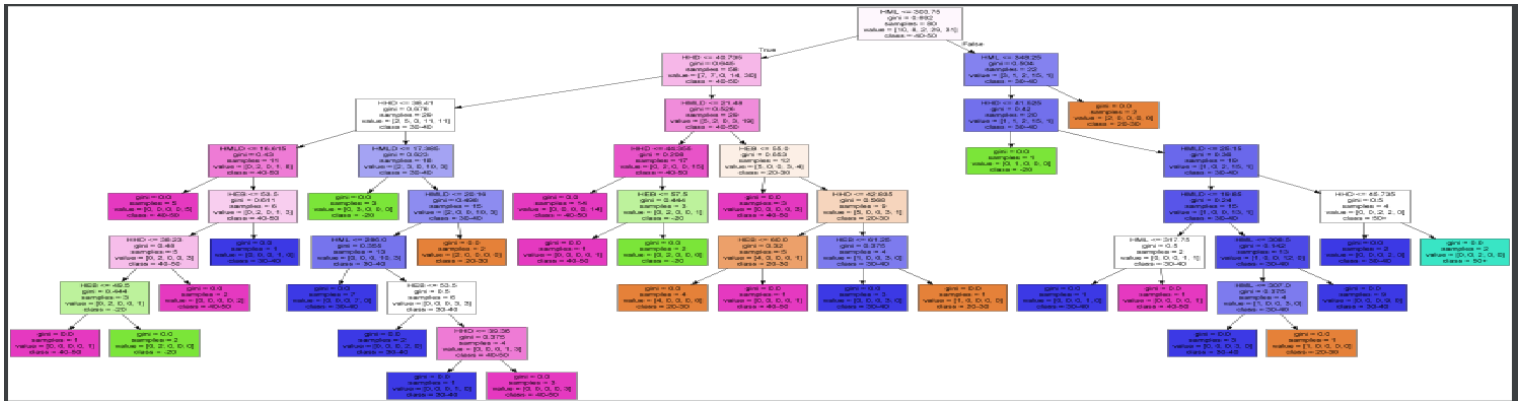

Figura 2.8: Arbore generat pentru un set de date cu 100 de exemple, clasificare varsta pentru humerus.

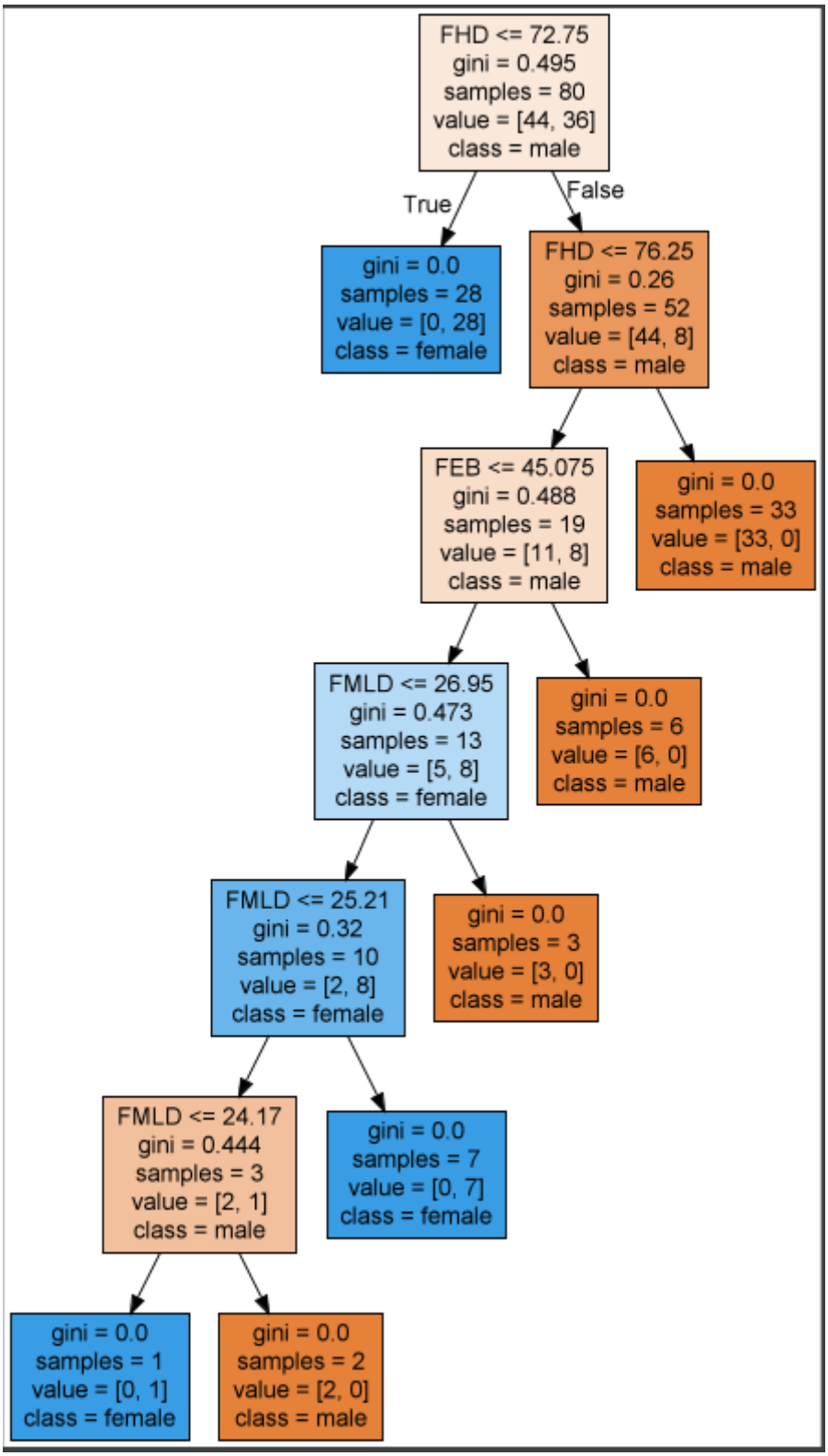

Figura 2.9: Arbore generat pentru un set de date cu 100 de exemple, clasificare sex pentru femur.

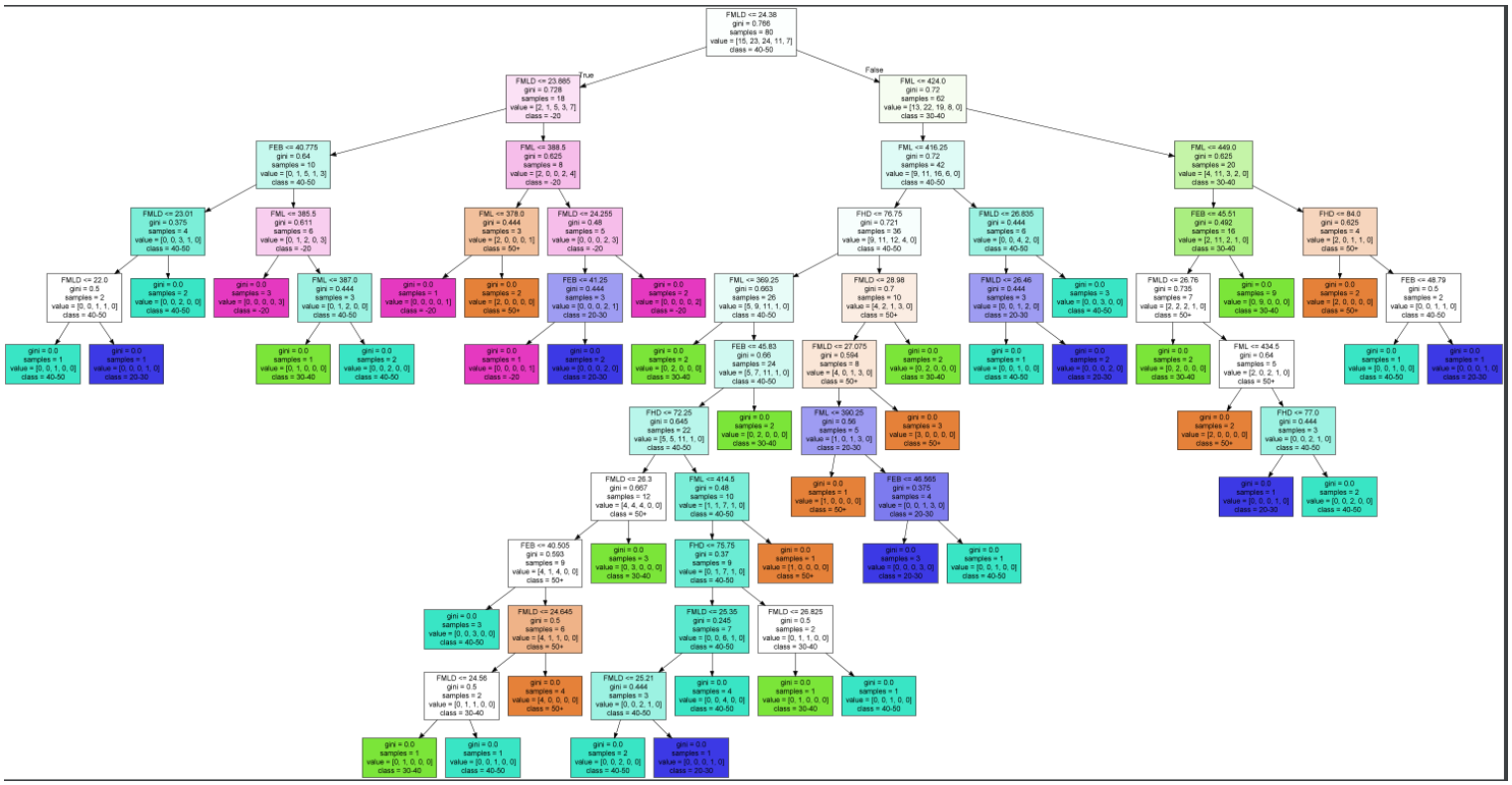

Figura 2.10: Arbore generat pentru un set de date cu 100 de exemple, clasificare varsta pentru femur.

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).

3.2 Retele neuronale

Lucrare stiintifica care sta la baza alegerii de a folosi un ANN pentru a determina sexul este *Determination of Gender from Pelvic Bones and Patella in Forensic Anthropology: A Comparison of Classification Techniquees.*

In acesta lucrare s-a folosit un algoritm de Back propagation Neural Network (BPNN) iar identificarea sexului unei prsoane se face pe baza oaselor perviene. Pentru antrenarea modelului s-au folosit 136 de oase pelviene (55 feminine si 81 masculine), unde 70% din ele au fost folosite pentru antrenarea modelului si 30% pentru testare. Modelul a fost antrenat pe baza inailtime, latimii si grosimii osului pelvian.

Acuratetea optinuta pentru osul pelvian drept este de 98.5% (datele de antrenament) si 98.3% (datele de testare) iar pentru osul pervian stang este de 98.49% (datele de testare) si 86.6% (datele de antrenament).

Asadar comparative cu rezultatele optinute la modelul implementat de noi care are o acuratete de 79.65%(datele de training), 87.58%(datele de testare) pentru humerus si 81.84(date de training), 82.93 (date de testare) pentru femur, modelul din lucrarea stiintifica prezentata da rezultate mult mai bune.

Capitolul 4

Imbunatatirea aplicatiei

4.1 Arbori de decizie

Comparand aplicatia cu varianta initiala se poate observa o imbunatatire a acuratetei, dar si a timpului de executie. Prelucrand setul de date a fost redus numarul de instante cu care se antreneaza algoritmul (pentru femur exista 300, iar pentru humerus 320) astfel incat datele sa fie distribuite egal din punctul de vedere al sexului, dar si al varstei. Cu aceasta modificare acuratetea pentru sex a crescut de la aproximativ 65% la aproximativ 85%, iar pentru varsta acuratetea a crescut de la aproximativ 30% la aproximativ 45%. De asemenea, reducerea numarului de instante cu care se lucreaza de la aproximativ 2000 la aproximativ 300 a imbuntatit si timpul de executie, determinarea sexului si varstei fiind mult mai rapide.Astfel, masurand timpul de executie avem 0.103 secunde pentru determinarea sexului si 0.000998 secunde pentru determinarea varstei fata de cel putin 1-2 secunde cat era initial (pentru aceste rezultate s-a folosit libraria time din Python, iar timpul a fost masurat ca diferenta intre momentul in care incepe antrenarea arborelui de decizie si momentul in care s-a terminat antrenarea si s-a prezis valoarea pentru instanta data).

4.2 Retele neuronale

Comparand aplicatia cu varianta initiala se poate observa o imbunatatire a acuratetei. Prelucrand setul de date, a fost redus numarul de instante cu care se antreneaza algoritmul (pentru femur exista 300, iar pentru humerus 320) astfel incat datele sa fie distribuite egal din punctul de vedere al sexului, am modificat learning rate-ul astfel incat acesta sa scada pe parcurs ce modelul se antreneaza si a crescut numarul de epoci de la 200 la 800. Cu aceste modificari acuratetea pentru detectare de sex a crescut de la 79.65%(datele de training), 87.58%(datele de testare) la 88,6%(datele de training), 96,8%(datele

de testare) pentru humerus si de la 81.84%(date de training), 82.93% (date de testare) la 90.83%(date de training), 93.33% (date de testare) pentru femur.

4.3 Tehnici de evaluare

Functia de loss este functia care calculeaza distanta dintre iesirea curenta a algoritmului si iesirea asteptata. Este o metoda de a evalua modul in care algoritmul modeleaza datele. (vezi figura 4.5 si 4.6)

O matrice de confuzie este un rezumat al rezultatelor predictiei pentru o problema de clasificare. Numarul de predictii corecte si incorecte este rezumat cu valori de numarare si impartit pe fiecare clasa. Aceasta este cheia matricei de confuzie. Matricea de confuzie arata modalitatile in care modelul de clasificare este confuz cand face predictii. (vezi figura 4.7 si 4.8)

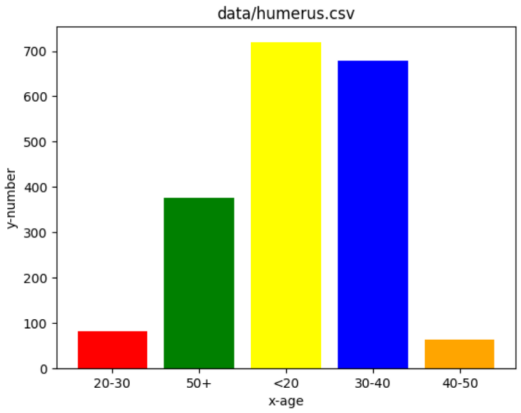

Figura 4.1: Distributie varsta initiala.

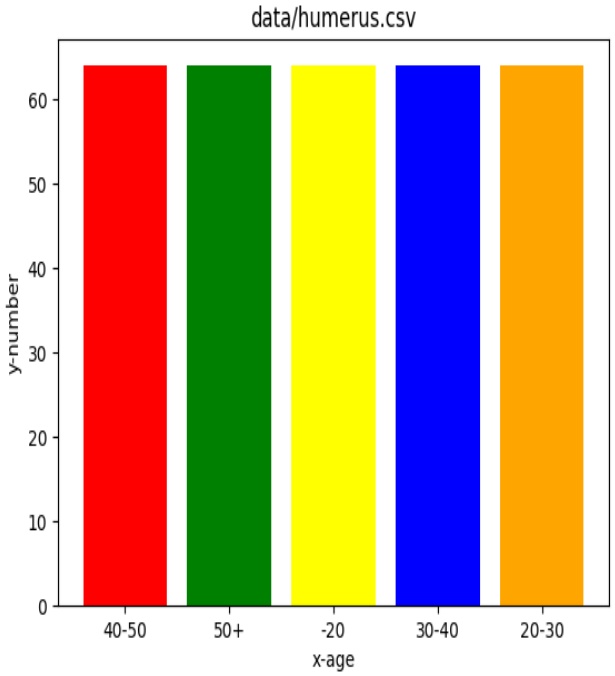

Figura 4.2: Distributie varsta actuala.

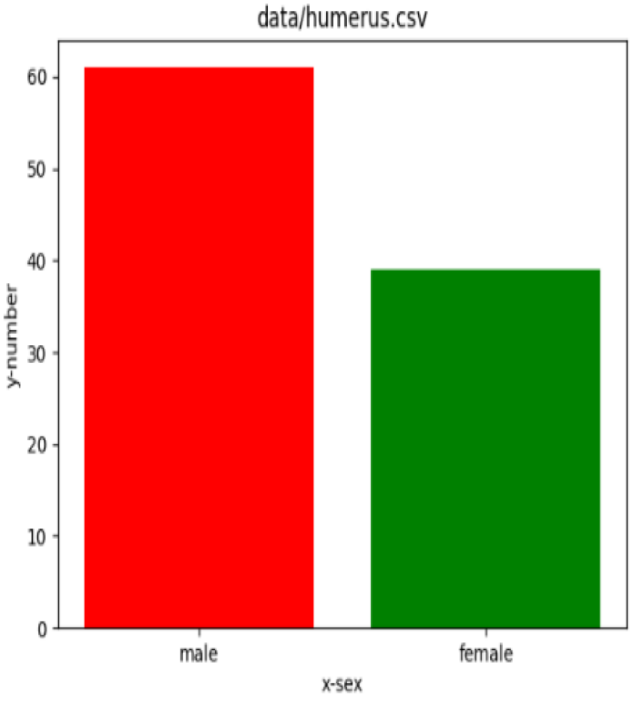

Figura 4.3: Distributie sex initiala.

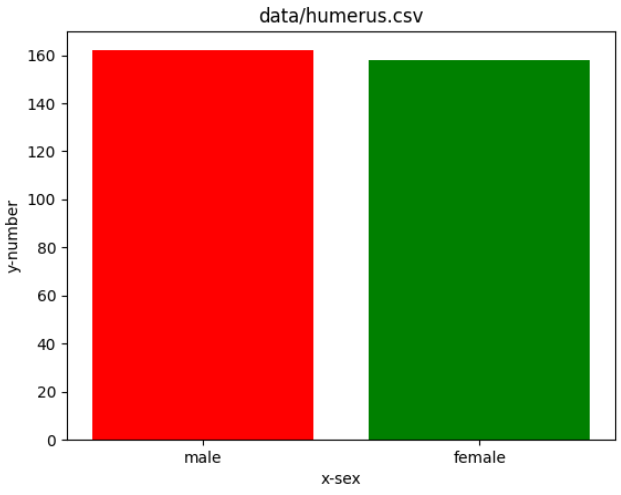

Figura 4.4: Distributie sex actuala.

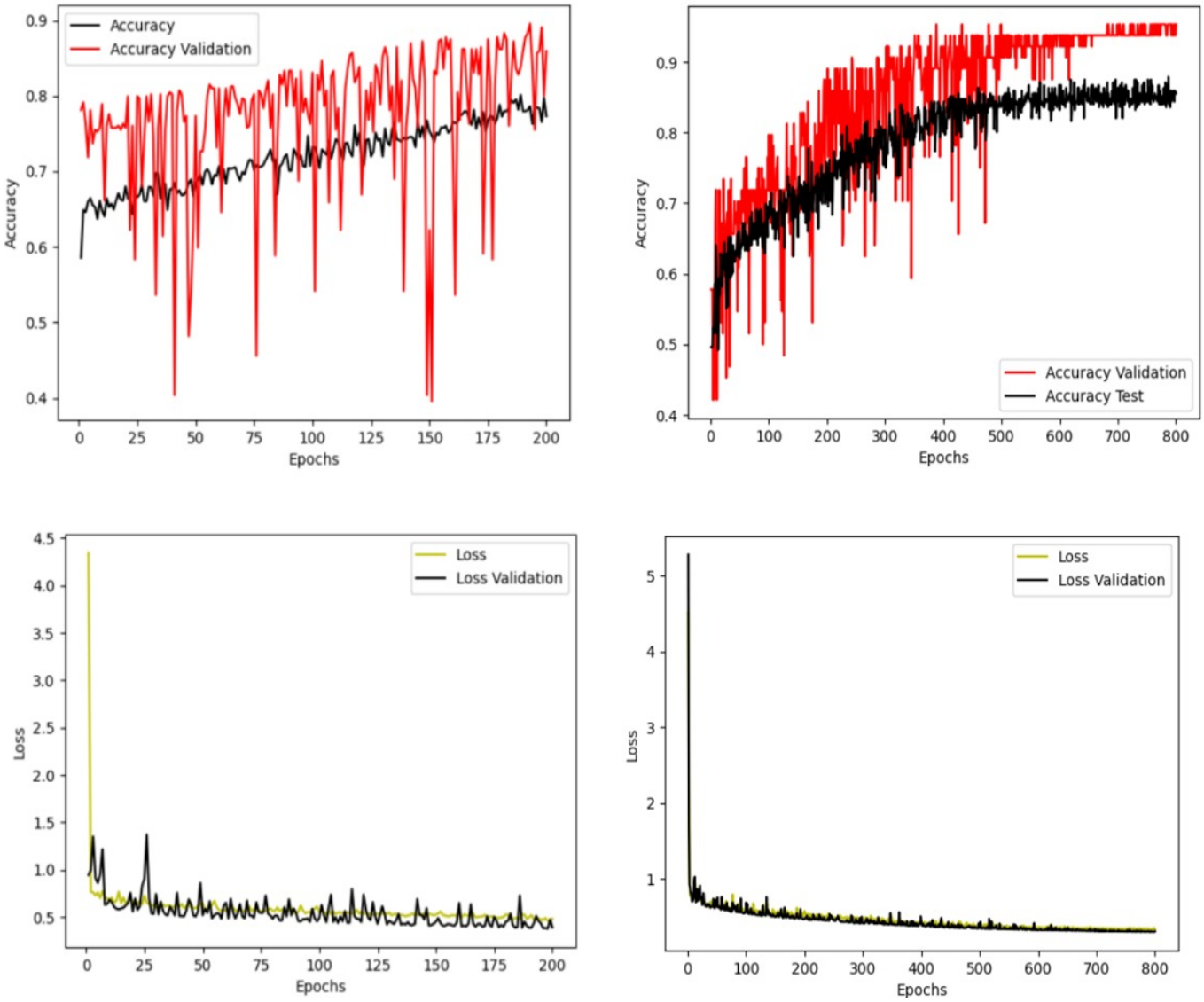

Figura 4.5: Acuratetea si loss-ul initial si actual pentru femur.

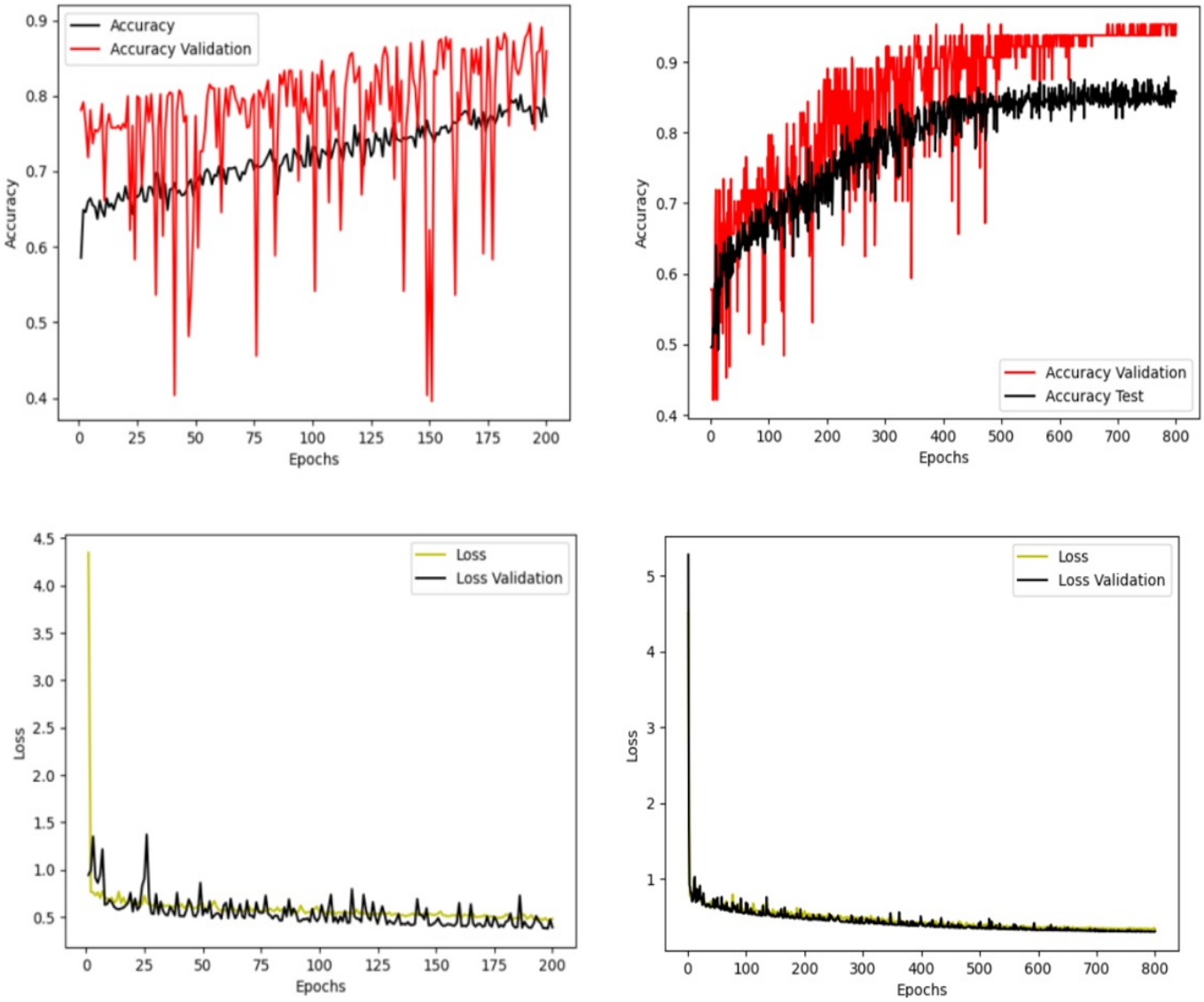

Figura 4.6: Acuratetea si loss-ul initial si actual pentru humerus.

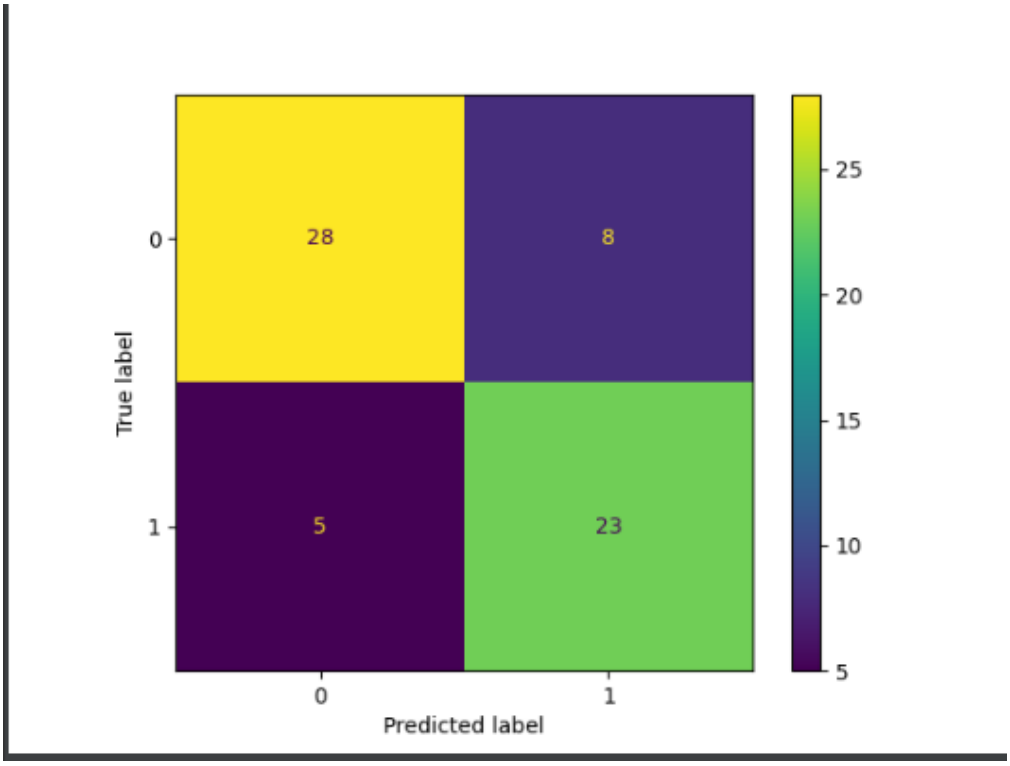

Figura 4.7: Matricea de confuzie sex humerus.

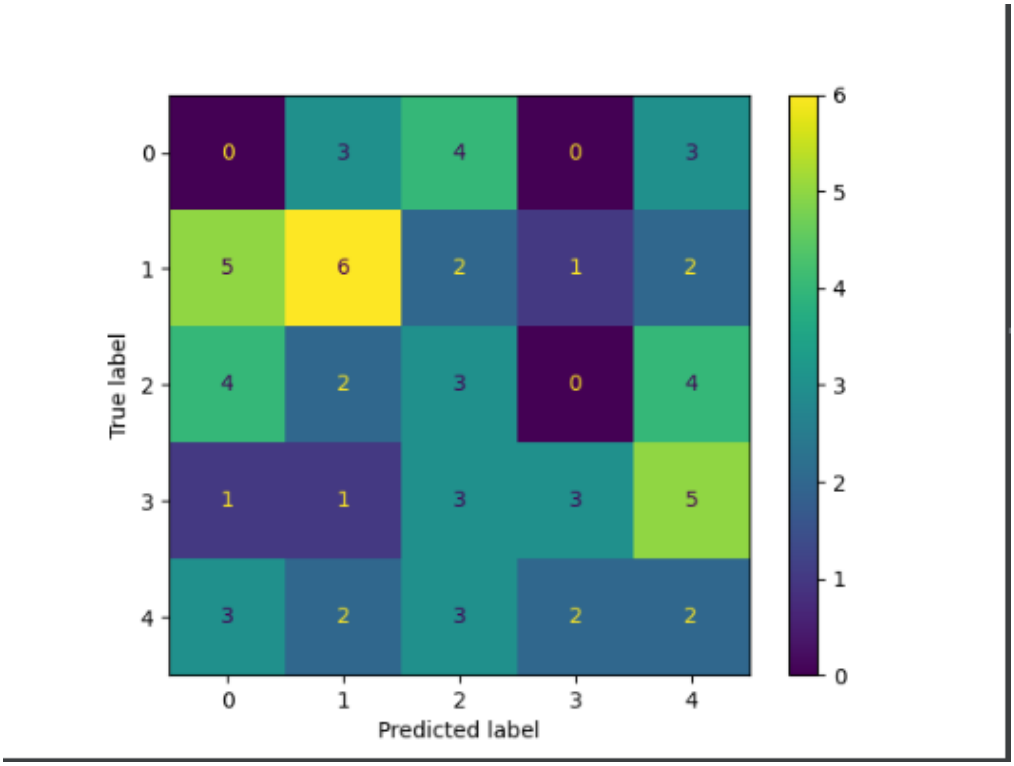

Figura 4.8: Matricea de confuzie varsta humerus.

Capitolul 5

Analiza statistica a algoritmilor

S-a realizat o analiza pe diferite seturi de date atat pentru humerus, cat si pentru femur urmarind acuratetea rezultata pentru determinarea sexului, dar si pentru determinarea varstei. Astfel, pentru fiecare tip de os s-au folosit cinci seturi de date cu diferite procente folosite la extragerea datelor de test din tot setul de date. Astfel, s-au folosit urmatoarele valori pentru a imparti datele: 20%, 40%, 80%, 35% si 70%.

5.1 Arbori de decizie (vezi 5.1)

Determinare sex pentru humerus

- media acuratetilor este: 81.41
- interval de incredere: [78.12, 84.37]

Determinare varsta pentru humerus

- media acuratetilor este: 28.73
- interval de incredere: [20.31, 33.59]

Determinare sex pentru femur

- media acuratetilor este: 87.08
- interval de incredere: [84.76, 88.75]

Determinare varsta pentru femur

- media acuratetilor este: 29.59
- interval de incredere: [23.33, 41.66]

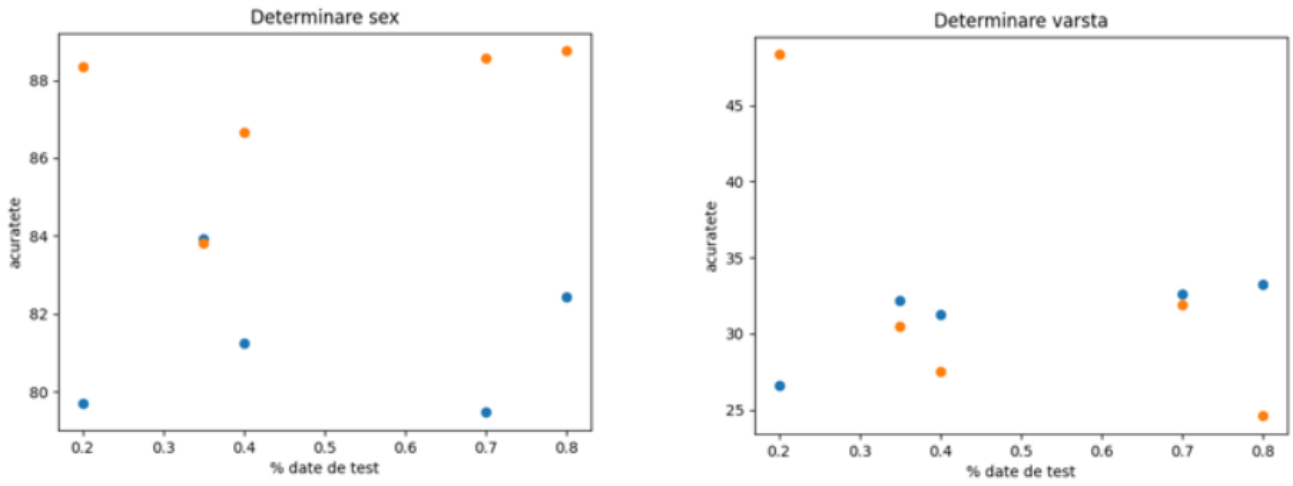

Figura 5.1: Reprezentare grafica a relatiei dintre acuratete si dimensiunea setului de date. (albastru - humerus, portocaliu - femur)

5.2 Retele neuronale

5.3 Rezultatele analizei

Capitolul 6

Wiki

Sklearn

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

Keras

- Structura retelei neuronale
- Straturile retelei neuronale

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 7

Lucrari stiintifice

Punctul de start al proiectului dat a fost constituit de urmatoarele lucrari stiintifice: [2], [5], [6], [4], [1], [7], [3], [8]. 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.

[8] M. GÃ¼lhal BOZKIR S. Deniz AKMAN, Pinar KARAKAfi. *The Morphometric Measurements of Humerus Segments.* 2005.