**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI, FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**



# *Dezvoltare Software pentru analiza datelor*

*Analiza componentelor principale*

*~animalele de la Grădina Zoologică~*

**Coordonator științific: Student:**

**Proșcanu Cosmin-Adrian Țone Iulia-Paula**

**grupa 1093, seria E**

**București, 2022**

Cuprins

[**Motivarea alegerii temei și a metodei de analiză** 3](#_Toc92222611)

[**Setul de date ales** 3](#_Toc92222612)

[**Rezultate și interpretări** 4](#_Toc92222613)

[**Concluzie** 10](#_Toc92222614)

# **Motivarea alegerii temei și a metodei de analiză**

În cadrul acestui proiect realizat pentru disciplina Dezvoltare Software pentru analiza datelor, am ales să efectuez analiza în componente principale pentru setul de date menționat anterior, din diverse motive pe care le voi expune mai jos.

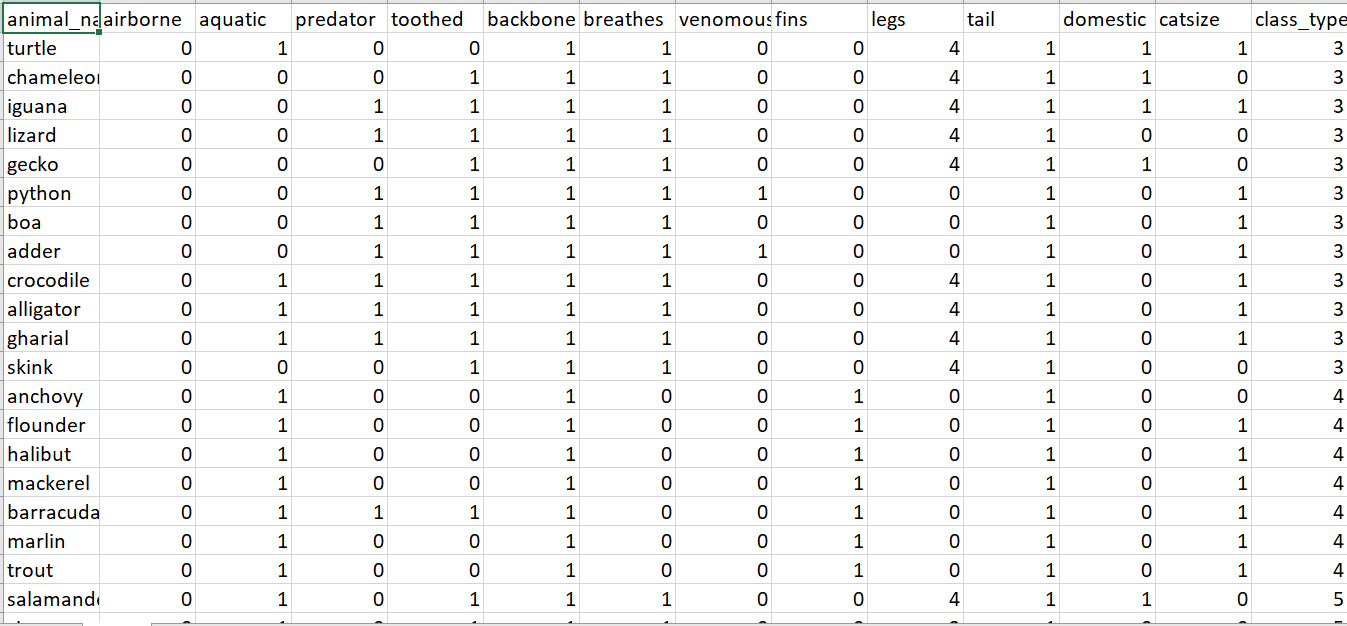
Spre de deosebire de analiza factorială – unde se modelează variaţiile variabilelor X prin transformări liniare a unui număr fixat, limitat de factori numiţi ”ascunşi”, latenţi - ACP caută combinaţii liniare între variabile, ordonându-le după valorile proprii ale matricei covarianţelor.

Analiza în Componente Principale permite abordarea caracterului multidimensional a datelor (variabilelor) ce caracterizează un individ. Principiul fundamental al acestei metode este de a extrage cel mai mic număr de componente care să recupereze cât mai mult din informaţia totală conţinută în datele originale, aceste noi componente exprimând atribute noi ale indivizilor şi construite astfel încât să fie necorelate între ele, fiecare din aceste noi variabile fiind o combinaţie liniară de variabile originale.

Deşi la baza acestei metode stă acelaşi principiu ca şi la analiza factorială (în principiu este o metodă factorială liniară), analiza în componente principale diferă de aceasta prin modul de definire a elementelor tabelului de date iniţial şi modul de calcul al distanţelor dintre puncte. Ca metodă descriptivă de analiză a datelor se aplică doar variabilelor cantitative şi tabelelor de dimensiuni mari care cuprind informaţii referitoare la mai mult de 15 indivizi şi 4 variabile. De aceea, am ales-o pentru a analiza influența anumitor trăsături ale animalelor de la zoo.

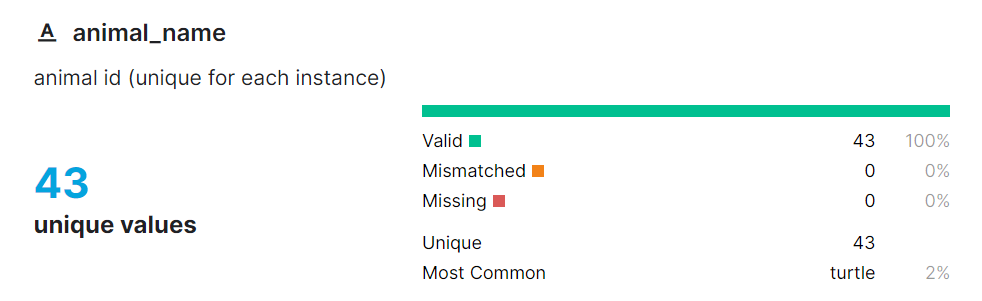
# **Setul de date ales**

În acest set de date, denumit *zoo2.csv* și extras de la adresa <https://www.kaggle.com/agajorte/zoo-animals-extended-dataset> , se află 43 de observații care conțin detalii despre 13 trăsături, precum în imaginea de mai jos.

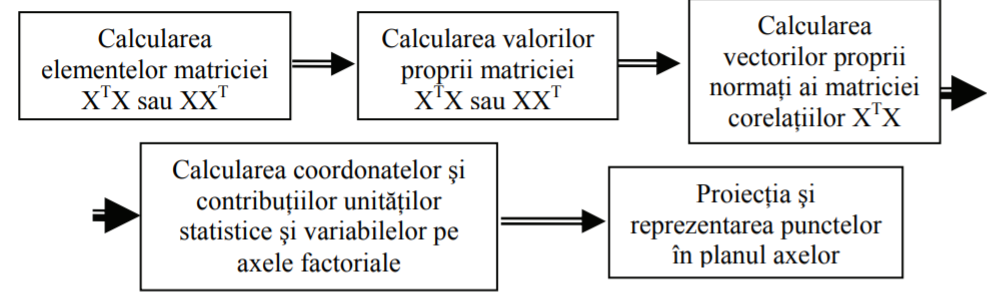


Datele sunt extrase în urmă cu 3 ani, precum este specificat și pe site și reprezintă animale din diferite clase, exceptând Mamiferele și Păsările, fiind extrem de întâlnite în multe alte seturi de date. Acest set de date este des utilizat în Machine Learning, și astfel l-am ales pentru a obține rezultate concludente și ușor de interpretat. De asemenea, este indicat pentru a fi utilizat în clasificări.

Valorile sunt unice și valide.

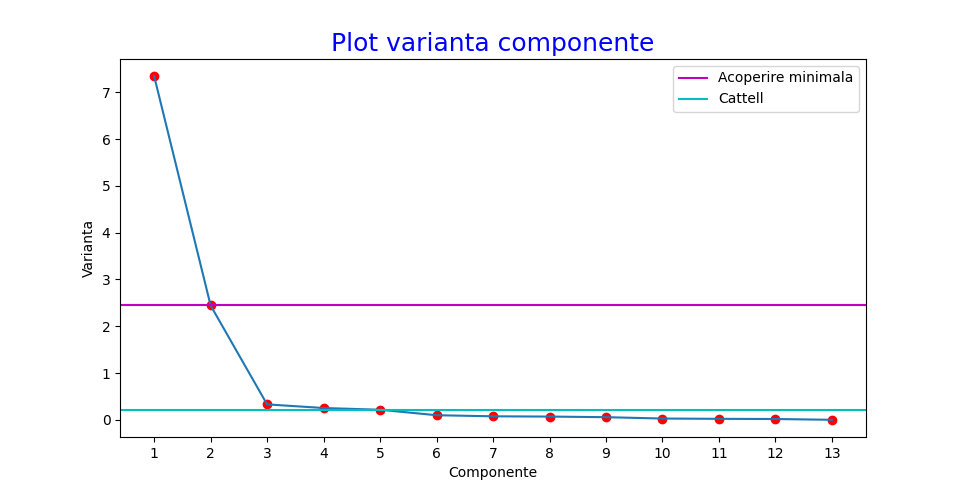


# **Rezultate si interpretari**

 Etapele urmate în Analiza Componentelor Principale pot fi sintetizate în graficul de mai jos. Acești pași îi voi urma și în codul atașat proiectului.

Primii pași în conceperea unei analize sunt constituiți de pregătirea unui script corespunzător (precum cel de la seminar), care conține fișiere .py cu funcții, grafice și clase concepute în acest sens, toate integrate și apelate în fișierul main.

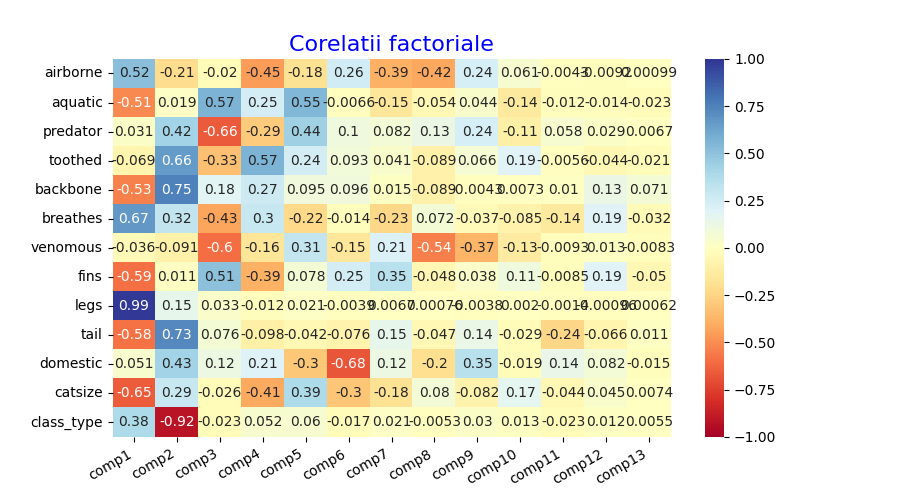
Am importat librăriile necesare, am extras datele din fișierul csv, eliminând valorile NaN și păstrând totodată într-un vector numele variabilelor observate pentru utilizarea lor ulterioară, dar am și creat modelul ACP din clasa eponimă. Funcțiile integrate au generat și grafice.

Astfel, în graficul anterior creat se află varianța componentelor din cadrul modelului meu. Componentele principale sunt construite sub formă de combinații liniare de variabile inițiale, care concentrează o cât mai mare parte din varianță. Fiecare componentă preia o parte din varianța ramasă neexplicată de cele precedente.

O imagine care conține text

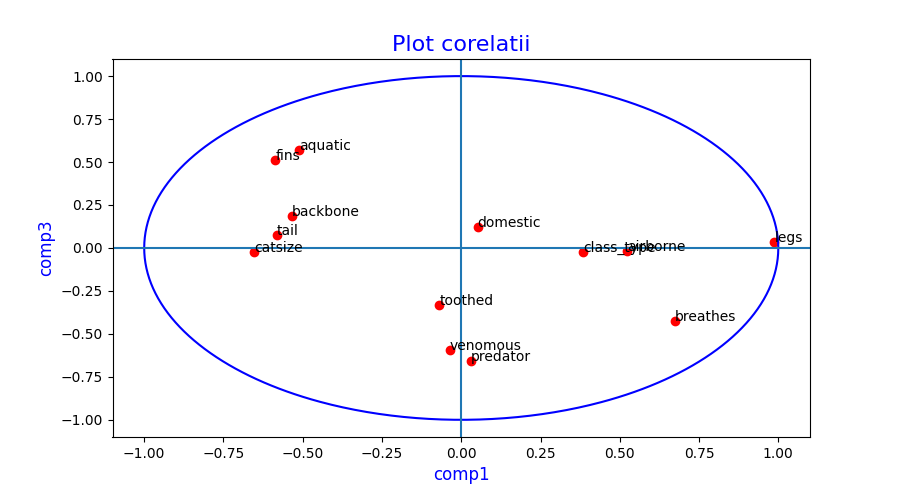
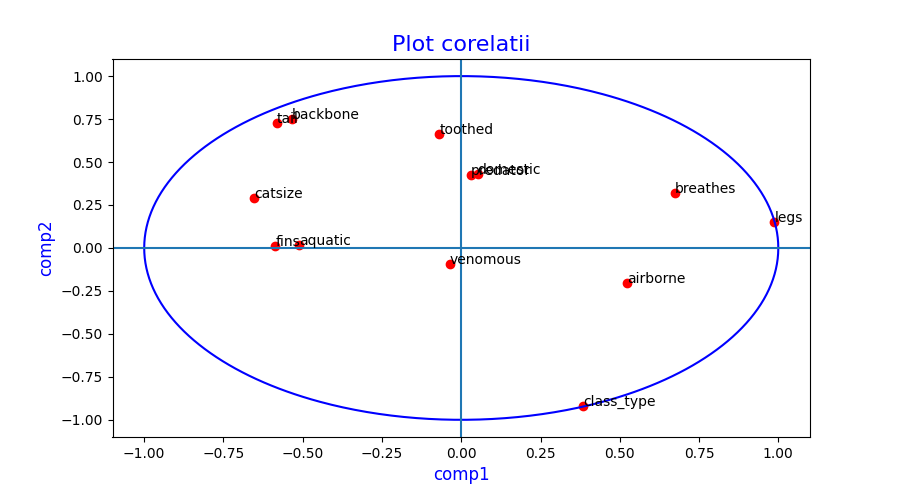
Descriere generată automatApelul funcției de varianță generează și un fișier csv:

Astfel, sunt salvate date pentru fiecare componentă. Varianța pentru componenta 1 este 7.34, înseamnând că aceasta concentreză de 7.34 ori mai multă informație decât o variabilă obișnuită. Aceasta explică o mare parte din varianța totală, mai exact 66%. De asemenea, este reprezentată și varianța cumulată, dar și procentul cumulat. Echivalent se interpretează și datele pentru celelalte componente.

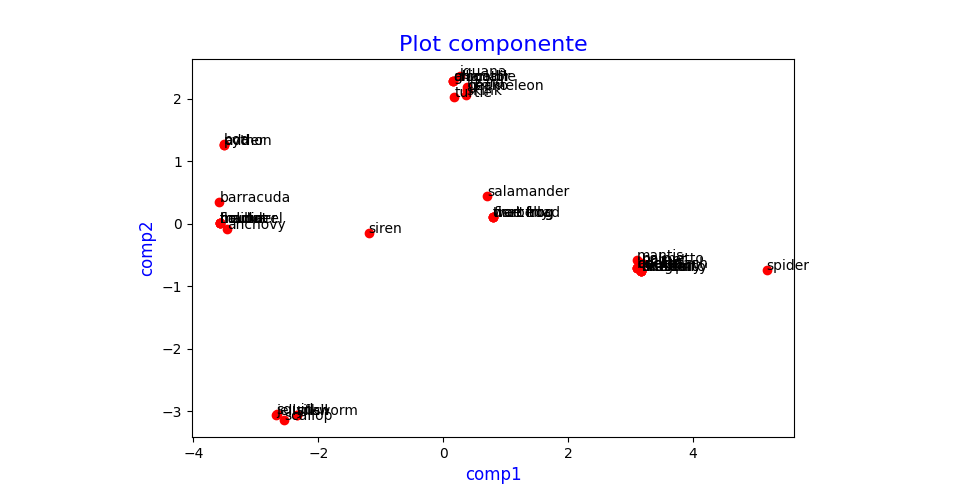
În continuare, urmează corelațiile factoriale, cu cât o valoare este mai apropiată în modul de 1, cu atât acea variabilă este mai semnificativă pentru componenta aleasă. Observăm că

numărul de picioare are cea mai mare corelație, de 99% pentru componenta 1, care reprezintă airborne – viața aeriană a animalului. Astfel, ne așteptăm ca variabilele cu mai multe picioare să fie în dreapta plotului instanțelor care au componenta 1 pe axa OX. Astfel, *spider*, care are cele mai multe picioare, este foarte în dreapta pe graficul componentelor de la pagina 7.

Aceste colerații pot fi vizualizate într-o manieră extrem de facilă și în graficele de mai jos, în care pe axele Ox și Oy avem primele două componente, pentru primul grafic, și componenta 1 cu 3 în cel de-al doilea, deoarece în ele regăsim cele mai mari valori.



De asemenea, s-a creat și un grafic, cel de mai jos, în care sunt reprezentate toate cele 43 de observații inițiale, animale, pe fiecare componentă. Astfel, se observă coeficientul pe scorurile, adică componentele principale standardizate:



O imagine care conține text

Descriere generată automatUlterior, se calculează cosinusurile, care denotă faptul că o instanță este cu atât mai bine reprezentată pe o axă cu cât cosinusul unghiului format între vectorul punct asociat instanței și axa respectivă este mai mare:

Fiind un număr relativ mare de componente, nu au încăput toate valorile pe ecran. Se observă că unele componente au un cosinus de aproape 1, precum animalul *gharial[[1]](#footnote-2)*pentru cea de-a doua componentă, denumită aquatic, ce arată dacă animalul respectiv provine din mediul acvatic sau nu. Acest lucru se justifică, gharialul fiind un prădător acvatic, reptilă asemănătoare crocodililor și aligatorilor.

O imagine care conține reptilă, exterior, reptilă crocodiliană, apă

Descriere generată automat

În continuare, se află contribuțiile, care arată ce procent din varianța axei este datorat instanței alese. O parte din aceste contribuții salvate și într-un csv le-am ilustrat mai jos :

O imagine care conține text

Descriere generată automat

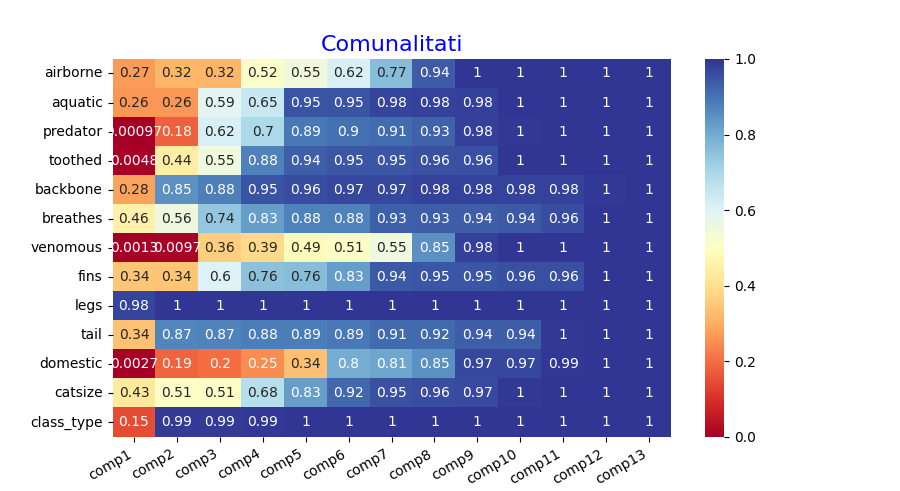
Din nou, se observă o contribuție mare a peștilor și a altor animale marine pentru componenta cu numărul doi, care specifică acvaticitatea animalelor, fapt ce se justifică de altfel. Am observat un procent uriaș, de 23% pentru broasca țestoasă și componenta 6 – respiră. Fiind multe animale marine în set, deci care nu respiră prin plămâni, broasca țestoasă prezintă o contribuție majoră.

Urmează comunalitățile - care reprezintă cantitatea de varianță explicată în comun de către un grup din componentele principale. O parte din valorile calculate sunt ilustrate și mai jos.

O imagine care conține text

Descriere generată automat

Pentru o mai bună vizualizare a rezultatelor, voi atașa și graficul desenat.



Aceasta manifestă procentul varianței unei variabile explicată de factorii reuniți, și poate fi interpretată ca siguranța indicatorului reprezentat de acea variabilă.

Valorile mici indică faptul că o caracteristică nu este bine reprezentată de acea componentă, astfel de valori regăsindu-se pentru proprietatea *venomous.* În schimb, *legs* are cele mai mari valori.

# **Concluzii**

În urma analizei efectuate, am obținut o sumedenie de rezultate asupra celor 43 animale de la zoo, ce reprezintă observații, și cele 13 variabile inițiale, proprietăți ale acestor viețuitoare necuvântătoare, constituite în componente. Prin puterea acestei metode de analiză a datelor, am realizat diferite grafice rezultate în urma funcțiilor modelului ACP, pe care le-am intepretat anterior.

Componentele ACP explică maximul varianței și sunt ortogonale între ele, fiind combinații liniare are variabilelor observate. Această metodă este una observațională de reducere a dimensionalității.

1. vezi <https://en.wikipedia.org/wiki/Gharial> [↑](#footnote-ref-2)