Proiect cercetare Lab3

**Schita cuprins:**

Abstract: Rezumat despre scopul lucrarii, contur al obiectivelor cheie: analiza statistica in R a diferitelor date despre oua de pasari precum descoperirea corelatiilor, outliers si identificarea de tipare asupra diferitelor variabile (culoare, transparenta, specie, dimensiuni, etc.). Rezumat al metodologiilor folosite: analiza multivariata, clustering, clasificare. Motivul abordarii analizei in R.

Introducere: Importanta studiului pentru descoperirea adaptarilor speciilor de pasari si impactul mediului precum si obtinerea de date ce pot fi interpretate de specialisti in domeniu. Obiectivul de a descoperi tipare ce nu se pot vedea cu ochiul liber, modelarea si generalizarea unui proces de studiu al datelor din domeniul biologic.

Lucrari conexe: Diferite metodologii de clustering prezentate si elaborate in diferite lucrari ([[15]](#Bookmark1), [[13]](#Bookmark1), [[12]](#Bookmark1), [[1]](#Bookmark1), [[2]](#Bookmark1), [[3]](#Bookmark1), [[4]](#Bookmark1)), analiza regresies [[5]](#Bookmark1), metode statistice ([[6]](#Bookmark1), [[7]](#Bookmark1), [[10]](#Bookmark1)), biostatistica in R ([[8]](#Bookmark1), [[9]](#Bookmark1)), statistica in R ([[14]](#Bookmark1),[[11]](#Bookmark1)).

Metodologie: Prezentarea seturilor de date si a variabilelor prezente (specie, procentaje de lumina ce patrund prin coaja oului, dimensiunile oualelor, habitat, culoarea oualelor, timpul de incubare, etc.). Identificare si sustragerea de outliers urmat de vizualizarea distributiilor si a variabilitatii prin boxplots, pairwise scatterplots. Aplicarea componentei principale pentru reducerea dimensiunilor si prezentarea corelatiilor gasite prin matrice de corelatie. Introducerea cluster analysis si evaluarea acestora prin silhouette analysis si Hubert Index precum si vizualizare. Prezentarea arborilor de clasificare obtinuti.

Rezultate: Explicarea distributiilor si corelatiilor calculate dintre variabile. Identificarea datelor ce contribuie variabilitatii. Impactul outilers asupra seturilor de date. Examinarea clusterilor obtinute si a numarului optim de clustere.

Limitatii: Limitele impuse de seturile de date precum selectia prezenta si numarul de intrari.

Concluzie: Rezumat al constatarilor cheie si importanta acestora, potentiale cercetari viitoare si posibile aplicatii si interpretari viitoare.

**Ciclu de viata:**

Planificarea lucrarii: Stabilirea de obiective si scopul lucrarii, obtinerea datelor. Organizarea acestui document Word.

Pre-procesare: Diferite selectii asupra seturilor de date si manipularea valorilor lipsa, identificarea si sustragerea de outliers precum si normalizarea valorilor.

Exploratory data analysis: Vizualizarea distributiilor si relatiilor initiale prin boxplot/scatterplots. Calculul matricei de corelatie.

Analiza multivariata: Analiza de componenta principala, aplicarea k-means clustering si analiza si optimizarea clusterelor formate si vizualizare.

Clasificare: Constructia, interpretarea si vizualizarea arborelui de decizie.

Procesare rezultate si vizualizare de ansamblu.

Documentatie si prezentare metode folosite in R. Organizarea programului pentru a permite analiza exterioara si reproducere.

**Functionalitati:**

Analiza multivariata, analiza de componenta principala;

Corelatii si identificare outliers;

Metode de clustering si determinarea parametrilor si numarului optim de clustere;

Clasificare si arbori de decizie;

Vizualizarea datelor;

Patrundere in procesul tehnic (prezentare fragmente de cod).

**Bibliografie:**

[1] Shi Na, Liu Xumin and Guan Yong, “Research on k-means Clustering Algorithm: An

Improved k-means Clustering Algorithm”, Proc. ㅤ of the 3rd International Symposium

on Intelligent Information Technology and Security Informatics, 2010.

[2] Zahra Nazari, Dongshik Kang, M. Reza Asharif, Yulwan Sung and, Seiji Ogawa, “A new

hierarchical clustering algorithm”, Proc. ㅤof the International Conference on Intelligent

Informatics and Biomedical Sciences (ICIIBMS), 2015.

[3] Nisha and Puneet Jai Kaur, “Cluster quality based performance evaluation of hierarchical

clustering method”, Proc. ㅤof the 1st International Conference on Next Generation

Computing Technologies (NGCT), 2015.

[4] Jolliffe IT, Cadima J. 2016 Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A 374: 20150202. <http://dx.doi.org/10.1098/rsta.2015.0202>.

[5] Gülden Kaya Uyanık and Neşe Güler, “A Study on Multiple Linear Regression Analysis”,Proc. of the 4th International Conference on New Horizons in Education, 2013.

[6] Baron, Michael. 2019. Probability and Statistics for Computer Science. 3rd ed. Boca Raton, FL: CRC Press, Taylor & Francis Group.

[7] Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer. 2008. Mathematical Statistics with Applications. 7th ed. Belmont, CA: Thomson Brooks/Cole.

[8] Källén, Anders. 2011. Understanding Biostatistics. Chichester, UK: John Wiley & Sons.

[9] Shahbaba, Babak. 2012. Biostatistics with R: An Introduction to Statistics Through Biological Data. New York: Springer.

[10] Härdle, Wolfgang Karl, and Léopold Simar. 2019. Applied Multivariate Statistical Analysis. 5th ed. Switzerland: Springer Nature.

[11] Dalgaard, Peter. 2008. Introductory Statistics with R. 2nd ed. New York: Springer Science+Business Media.

[12] K.Mandeep, K.Usvir, Comparison Between K-mean and Hierarchical Algorithm Using Query Redirection, IJARCSSE, Jul. 2013.

[13] S.Chatti, G.R. Krishna et.al, A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set, Procedia Computer Science.

[14] P. Cichosz, Data Mining Algorithms Explained Using R, John Wiley & Sons, Ltd. 2015.

[15] Arpit Gupta, Ankit. Gupta & A. Mishra, “Research paper on clustering techniques of data variations”, International Journal of Advance Technology & Engineering Research (IJATER), vol. 1, issue. 1, November 2011.