Análisis de Datos I

Primera fase de proyecto

Moisés Monge Cordonero - Joshua Cervantes Artavia

Descripción general de los datos

Los datos fueron obtenidos del repositorio de Machine Learning de la Universidad de California Irving, la cual se puede accesar en el siguiente link: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease y este dataset tiene por nombre Chronic_Kidney_Disease el cual contiene una serie de variables relacionadas a 400 pacientes de los cuales 250 presentan enfermedad del riñón y 150 no. Esta data fue accesada el día 28 de Agosto del 2023. Los individuos en cuestión responden a pacientes de Apollo Hospitals en Karaikudi, Taminaldu, India, durante un proceso de recolección de datos de al alrededor de 2 meses. Los datos fueron recogidos a partir de Julio del año 2015.

Variables

El dataset se compone por total de 25 variables de las cuales 11 son numéricas y 14 nominales. Las variables en cuestión son:

- age: Edad en años (numérica)
- bp: presión sanguínea en mm/Hg (numérica)
- sg: gravedad específica con valores entre (1.005, 1.01,1.015,1.020.1.025) (nominal)
- al: albúmina con valores en (0,1,2,3,4,5) (nominal)
- su: azúcar con valores en (0,1,2,3,4,5) (nominal)
- rbc: células de sangre roja con valores en (normal, anormal) (nominal)
- pc: células pus con valores en (normal, anormal) (nominal)
- pcc: grupos de cédulas de pus con valores en (presente, no presente) (nominal)
- ba: bacteria con valores en (presente, no presente) (nominal)
- bgr: glucosa en sangre aleatoria en mgs/dl (numérica)
- bu: urea en sangre en mgs/dl (numérica)
- sc: creatinina sérica en mgs/dl (numérica)
- sod: sodio en mEq/L (numérica)
- pot: potasio en mEq/L (numérica)
- hemo: hemoglobina en gms (numérica)

- pcv: hematocrito en % (numérica)
- wc: conteo de células blancas en cells/cumm (numérica)
- rc: conteo de células rojas en millones/cmm (numérica)
- htn: hipertensión con valores en (sí, no) (nominal)
- dm: diabetes mellitus con valores en (sí, no) (nominal)
- cad: enfermedad de arteria coronaria con valores en (sí, no) (nominal)
- appet: apetito con valores en (bueno, pobre) (nominal)
- pe: edema del pie con valores en (sí, no) (nominal)
- ane: anemia con valores en (sí, no) (nominal)
- class: si presenta o no enfermedad crónica del riñón -ckd-con valores en (ckd, notckd)

A las variables mencionadas en cuestión, se le han añadido cuatro variables relativas adicionales que se han construido, las cuales corresponden a las relaciones: hemo/wbcc, hemo/rbcc, bu/bgr, sod/bp. Como variable complementaria se considerará la variable class.

Data faltante

La data presenta algunos datos faltantes, para esto primeramente se han contabilizado aquellas observaciones que tuvieran 4 datos faltantes o más y se han excluido, dejando un total de 357 observaciones y posteriormente, sobre dichas 357 observaciones se ha realizado una imputación de datos por medio de un KNN.

Se ha escogido esta tabla de acuerdo al número de observaciones que posee, ya que con esta será posible visualizar de mejor forma los datos mediante un modelo de reducción de la dimensionalidad como puede ser un A.C.P.. En comparación con la tabla alternativa, empleada por Koklu y Ozkan (2020) en Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques, esta tabla posee una menor cantidad de datos 400 contra 13 611 datos que posee la alternativa. Además, la calidad de los datos de ambas tablas es buena.

Bibliografía

Koklu, M. and Ozkan, I.A., (2020), Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. *Computers and Electronics in Agriculture*, *174*, 105507.

Rubini,L., Soundarapandian,P., and Eswaran,P.. (2015). Chronic_Kidney_Disease. UCI Machine Learning Repository.