Nuria García Fuentes y Alicia Perdices Guerra

limpieza y análisis de datos

Práctica 2

## **Autoras:**

**Índice**

|  |  |
| --- | --- |
| CONTENIDO | Págs. |
| [Contexto](#Contexto) | 2 |
| [Dataset Estudio](#Dataset) | 2-3 |
| [Entrega GitHub](#GitHub) | 3-4 |
| [Referencias](#Referencias) | 4-5 |

**Contexto**

En el contexto de la práctica de Tratamiento, Limpieza y Visualización de datos, hemos querido seguir en la línea planteada desde el principio, investigando sobre los diversos patrones presentes en todo tipo de enfermedades. En este caso, el dataset de estudio versa sobre el diagnóstico de la Insuficiencia Renal Crónica partiendo de la información de analíticas y hábitos de pacientes del hospital Apollo en Tamil Nadu, en la India.

Se pretende descubrir las relaciones que puedan establecerse entre valores anómalos de variables como la Creatinina en Suero, sodio, potasio y la relación subyacente con la presencia de otras enfermedades como la Diabetes o Hipertensión.

**Dataset de estudio**

El dataset de estudio se encuentra disponible en: <https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease>.

Dicho dataset sigue la siguiente estructura:

* **Chronic Kidney Disease:**

|  |  |  |
| --- | --- | --- |
| Atributos | Descripción | Tipo de datos |
| Id | Identificador | Numerical |
| Age | age in years | Numerical |
| bp | Blood presure(bp in mm/Hg) | Numerical |
| sg | Specific Gravity(nominal).sg - (1.005,1.010,1.015,1.020,1.025) | String |
| al | Albumin(nominal):al - (0,1,2,3,4,5) | String |
| su | Sugar(nominal):su - (0,1,2,3,4,5) | String |
| rbc | Red Blood Cells(nominal):  rbc - (normal,abnormal) | String |
| pc | Pus Cell (nominal):pc - (normal,abnormal) | String |
| pcc | Pus Cell clumps(nominal):pcc - (present,notpresent) | String |
| ba | Bacteria(nominal):ba - (present,notpresent) | String |
| bgr | Blood Glucose Random(numerical):  bgr in mgs/dl | String |
| bu | Blood Urea(numerical):bu in mgs/dl | String |
| sc | Serum Creatinine(numerical):sc in mgs/dl | String |
| sod | Sodium(numerical):sod in mEq/L | String |
| pot | Potassium(numerical):pot in mEq/L | String |
| hemo | Hemoglobin(numerical:hemo in gms | String |
| pcv | Packed Cell Volume(numerical) | String |
| wc | White Blood Cell Count(numerical):  wc in cells/cumm | String |
| rc | Red Blood Cell Count(numerical):  rc in millions/cmm | String |
| htn | Hypertension(nominal):htn - (yes,no) | String |
| dm | Diabetes Mellitus(nominal):dm - (yes,no) | String |
| cad | Coronary Artery Disease(nominal):  cad - (yes,no) | String |
| appet | Appetite(nominal):appet - (good,poor) | String |
| pe | Pedal Edema(nominal):pe - (yes,no) | String |
| ane | Anemia(nominal):ane - (yes,no) | String |
| classification | Class (nominal):class - (ckd,notckd) | String |

**Entrega GitHub**

A continuación, detallamos los elementos de este estudio que se pueden encontrar en Github:

|  |  |
| --- | --- |
| Contribuciones | Firmas |
| Carpeta Kidney\_cleaning:   * kidney\_disease\_cleaning.Rmd:   (código R donde se realiza una limpieza previa de los datos. Imputación utilizada para valores nulos es la media.)   * kidney\_disease\_cleaning.html:   (Código en html generado a partir de kidney\_disease\_cleaning.Rmd, contiene la misma información pero en otro formato)   * kidney\_disease\_clean\_1.csv:   (dataset obtenido de la limpieza de kidney\_disease\_cleaning)   * kidney\_disease\_cleaning\_2.Rmd:   (código R donde se realiza una limpieza previa de los datos. Imputación utilizada para valores nulos knn valores más cercanos. Estudio de valores atípicos)   * kidney\_disease\_cleaning\_2.html:   (Código en html generado a partir de kidney\_disease\_cleaning\_2.Rmd, contiene la misma información pero en otro formato)   * kidney\_disease\_clean\_2.csv:   (dataset obtenido de la limpieza de kidney\_disease\_cleaning\_2. Dataset utilizado para el análisis de los datos)   * carpeta data:   Fichero origen desde donde comenzamos la limpieza: kidney\_disease.csv  Carpeta Kidney\_analysis:   * Analisis.Rmd:   (código R donde se realiza un análisis a partir del dataset limpio. Da respuesta a la práctica 2.)   * Analisis.html:   (Código en html generado a partir de Analisis.Rmd, contiene la misma información pero en otro formato)   * Analisis.pdf:   (Documento donde se encuentra la respuesta de la práctica. Corresponde al Análisis.Rmd y Analisis.html, pero en formato pdf)   * Carpeta data: kidney\_disease\_clean\_2.csv   (dataset obtenido en fase de cleaning y utilizado para el estudio) | Alicia Perdices  Núria García |
| Wiki:   * Wiki\_Limpieza y análisis\_Garcia\_Perdices\_Prac2.docx:   (documento donde se explica los ficheros entregados) | Alicia Perdices  Núria García |

**Referencias**

**Básicos**

* Calvo M., Subirats L., Pérez D. (2019). **Introducción a la limpieza y análisis de los datos**. Editorial UOC.

**Web**

* **Kaggle**. Datasets y análisis de éstos:

<https://www.kaggle.com/>

* **World kidney day**. Web sobre enfermedad crónica del riñón: <https://www.worldkidneyday.org/>
* El-Houssainy A.RadyaAyman S.Anwar. 2019**. Prediction of kidney disease stages using data mining algorithms**.sciencedirect:

<https://www.sciencedirect.com/science/article/pii/S2352914818302387>

* **Monte Yasur en Vanuatu: El volcán en activo más accesible del mundo.** Recuperado de:

<https://tiempodexplorar.com/monte-yasur-en-vanuatu-el-volcan-en-activo-mas-accesible-del-mundo/>

* Dr. Javier Muga Bustamante 06/09/2019 **Urea en la sangre alta:** <https://www.tuotromedico.com/parametros/urea-en-sangre-alta.htm#apartValores>
* Dr. Javier Muga Bustamante 03/09/2019 **Creatinina en la sangre alta:** <https://www.tuotromedico.com/parametros/creatinina-en-sangre-alta.htm#apartValores>
* Dr. Javier Muga Bustamante 07/11/2019 **Sodio en sangre bajo:**  [<https://www.tuotromedico.com/parametros/sodio-en-sangre-bajo.htm#apartValores>](https://www.tuotromedico.com/parametros/creatinina-en-sangre-alta.htm#apartValores)
* Dr. Javier Muga Bustamante 07/11/2019 **Potasio en sangre alto:**  [<https://www.tuotromedico.com/parametros/potasio-en-sangre-alto.htm#apartValores>](https://www.tuotromedico.com/parametros/creatinina-en-sangre-alta.htm#apartValores)
* 06/03/2018 **Hemoglobina en la sangre baja:**  [<https://www.tuotromedico.com/parametros/hemoglobina-en-sangre-baja.htm#apartValores>](https://www.tuotromedico.com/parametros/creatinina-en-sangre-alta.htm#apartValores)
* RnCeus.com**. “White Blood Cell Count (WBC) and Differential”**

<https://www.rnceus.com/cbc/cbcwbc.html>

* Mayo CLínic**, 03** Mayo, 2018 **“Síndrome diabético hiperosmolar”**

<https://www.mayoclinic.org/es-es/diseases-conditions/diabetic-hyperosmolar-syndrome/symptoms-causes/syc-20371501>

* **ggplot2 density plot : Quick start guide - R software and data visualization:** <http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization>
* **ggplot2 - Easy Way to Mix Multiple Graphs on The Same Page:**

<http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/#arrange-over-multiple-pages>

* Mohamed Helmy,November 29, 2016 “**R Boxplot Tutorial”:**

<https://rstudio-pubs-static.s3.amazonaws.com/231621_33654efbcf7b44da9103a3bc15406bec.html>

* Álvaro Alonso Fernández**,** 9 de abril de 2019. “**Gráficos en R”**

<https://rpubs.com/aafernandez1976/graficos>